PUBLIC SUBMISSION

SUNI Review Complete
Template=ADM-013
E-RIDS=ADM-03

ADD: Robert Roche-Rivera, Bridget Curran, Mary Neely
Comment (3)
Publication Date: 9/30/2021
Citation: 86 FR 54253

As of: 11/23/21 11:40 AM
Received: November 12, 2021
Status: Pending_Post
Tracking No. kvx-c9qm-b5m8
Comments Due: November 15, 2021
Submission Type: Web

Docket: NRC-2021-0166
Acceptability of ASME Section XI, Division 2, ‘Requirements for Reliability and Integrity Management (RIM) Programs for Nuclear Power Plants,’ for Non-Light Water Reactors

Comment On: NRC-2021-0166-0001
Acceptability of ASME Code Section XI, Division 2, Requirements for Reliability and Integrity Management (RIM) Programs for Nuclear Power Plants, for Non-Light Water Reactors

Document: NRC-2021-0166-DRAFT-0006
Comment on FR Doc # 2021-21295

Submitter Information

Name: Henry Stephens
Address:
Little River, SC, 29566
Email: hstephensjr@att.net
Phone: 17045196256

General Comment

See attached file(s)

Attachments

Response to Draft Regulatory Guide DG-1338 w Attachments

Attachment 2f-EPRI_3002010462_Human Factors in NDE__ A Literature Review and Field Observations_October 2021

HENRY M STEPHENS JR
hstephensjr@att.net
704-519-6256

To Whom It May Concern

United States Nuclear Regulatory Commission
Office of Adminstration
Progam Management

Subject: Comments to the Draft Regulatory Guide DG-1383 – Acceptability of ASME Code Section XI, Division 2, “Requirements for Reliability and Integrity Management (RIM) Programs for Nuclear Power Plants,” for Non-Light Water Reactors

As requested in the soliciation for public comments, please find herein (Attachment 1) detailed comments to the draft regulatory guide DG-1383 for the considerations to endorse ASME Section XI Division 2.

As a participant in the development and implementation of ASME Section XI including being a significant contributor in the development of ASME XI Division 2, RIM 2019 Edition, I am pleased to provide of suggestions and recommendations to DG-1383 for consideration by the USNRC.

I am currently retired from the EPRI where I worked for 34 years. While there, I served as the EPRI Prinicpal NDE Level III responsible for development and implementation of NDE Training and Qualification program. This included the Visual; UT IGSCC Detection, Sizing and Weld Overlay Qualifications (IE Bulletin 83-02); Section XI, Subsection IWE and IWL Containment Inspection Program; etc.

I have continued by active involvement in the nuclear power industry as a consultat and serving as the past President and Board Chair of the American Society for Nondestructive Testing (ASNT);Current Member of the ASNT Standards Development Committee that includes SNT-TC-1A and the ANSI/ASNT CP-189 Standard; past Chair of Section XI, Div. 1, WG Containment; Current Member of SG Water-Cooled Systems; Chair of Section XI, Div. 2, WG Monitoring and NDE (MANDE); and Member of SG Div. 2; and Chair of the ANDE Standards Committee.

If you have any questions or need addditional information, please contact me at your convenience.

Warm regards,

Henry M. Stephens, Jr.

Henry M. Stephens, Jr.
Consultant

Attachment

Attachment 1

(1) DG-1383 REFERENCE:

The Title, Applicability and many other areas contained in the Draft Regulatory Guide reference "Non-Light Water Reactors".

(1) BACKGROUND:

ASME XI Division 2 is applicable as a technology neutral standard and has been developed to be applied to both non-LWR as well as LWR technology. ASME XI, Division 2 permits the use of Division 1 criteria if it is appropriate for a given reactor design and distinguishes which PRA standard should be used in the development of a RIM program when addressing a non-LWR versus a LWR advanced reactor design.

It is understood that Title 10 CFR PART 50.55a establishes regulation for the use of ASME XI Division 1 as applied to light-water reactors. Further, ASME XI Division 1 addresses the existing fleet of PWR and BWR operating plant designs. Several advanced light-water reactors are currently in developments. The use of ASME XI Division 1 is neither adequate, nor appropriate for these newer designs. In some of these advanced LWR designs, the traditional safety cases that are a foundational consideration for the application of ASME XI Division 1, may not be as relevant as the safety cases used in the existing fleet design.

v

It is recognized that a license applicant for any advanced LWR design could request an alternative to 10CFR 50.55a but this may not be immediately obvious to some new reactor designers who may not have familiarity with the provisions of 10CFR50.55a (z).

(1) RECOMMENDATION:

It is recommended that DG-1383 be amended to provide guidance to users that the use of ASME XI Division 2 is a potential option for advanced LWR designs when ASME XI Division 1 is not applicable and directs them to the provisions of 10CFR50.55a (z).

(2) DG-1383 REFERENCE:

Basis for Regulatory Guidance Position 1 (page 6) and various other locations in the use of the word ***safety***. Specifically, "ARDC 1 states that SSCs important to safety are to be tested to quality standards commensurate with the importance of the **safety** functions to be performed."

(2) BACKGROUND:

The term safety is used throughout this draft RG. Reference to 10CFR50 Appendix B is also a cited. It is not clear if the term ***safety*** is meant to address traditional "Safety-Related' classifications (e.g., ASME Class 1, 2, 3 and Quality Groups A, B & C) or a broader use (e.g., worker/public safety). ASME XI Division 2 applies no specific safety classifications to SSC within a RIM Program. Instead, RIM requires that any SSC that is ***"risk significant"*** for the safe operation of a plant and therefore worker and public safety is to be an SSC within the scope of the RIM developed program. This might include SSC that would otherwise not be traditionally classified as "Safety-Related".

(2) RECOMMENDATION:

A description as to what is specifically intended by the use of the word “safety” would make it clearer as to what SSCs are specifically to be addressed.

(3) DG-1383 REFERENCE:

Basis for Regulatory Guidance Position 1 (page 7 - third bullet):
[For plants that do not have regularly scheduled refueling outages at frequencies of 5 years or less, proposed frequencies for the submission of owner’s activity reports (OARs) and owner’s repair/replacement certification (NIS-2) forms]

(3) BACKGROUND:

RIM establishes that an Inspection Interval shall not exceed 12 years (i.e., 144 months) irrespective of the time between refueling (e.g., on-line refueling vs traditional off-line refueling.) This was created with the recognition that some advanced reactor designs may not be taken out of service (e.g., “offline”) to accommodate a refueling outage. From an ASME Code perspective it would be at the end of this 12-year interval that an OAR-1 Form would be prepared.

It is understandable why the USNRC may not wish to wait until the end of a 12 interval to receive information contained in a prepared OAR-1 Form for twelve years and hence has established that a five-year periodicity for the preparation of an OAR-1 be created and submitted.

There are however two clarifications that would be useful to include in DG-1383 regarding this matter:

One is the descriptor of *refueling outage*. As noted, some advanced reactor designs may not have a traditional refueling outage, but instead a maintenance outage where the unit remains operational while refueling.

Two is that with the 2019 Edition of ASME XI, both in Division 1 and Division 2, the FORM NIS-2 is an attestation for the completion of a Repair/Replacement Activity (RRA).

The FORM NIS-2 does not include any significant or technical information. This change represents the incorporation of ASME Code Case N-532, where RRA that were required because an SSC had failed to an established ASME XI acceptance criteria would be documented on the OAR-1 Form and the FORM NIS-2 became abbreviated.

(3) RECOMMENDATIONS:

(1) Provide a clarification to the presently used term *refueling outage* to be more inclusive to evolutions such as “maintenance outages” or a description of a time frame for the desired submittal of an OAR-1 Form, exclusive of using the terminology refueling outage.
(2) Consider deleting the requirement to submit NIS-2 Forms.

(4) DG-1383 REFERENCE:

Basis for Regulatory Guidance Position 1 (page 7 – 4th bullet):
[“*Appropriate justification for flaw evaluation acceptance criteria for any components that exceed the temperature ranges in ASME Code, Section III, Division 1”.*]

(4) BACKGROUND:

ASME III Division 1 is for traditional LWR reactor designs and the normal operating temperature ranges of LWRs. In contrast, ASME III Division 5 is for High temperature reactors and includes provisions for certain HAA materials to operate in the creep regime with a stated maximum number of permissible cycles. This same limitation is found in RIM Appendix VII Article VII-3 .1. Consequently, it is not obvious why ASME Division 1 is cited for advanced reactor design flaw acceptance criteria.

(4) RECOMMENDATION:

Consider, revising this bullet so it is clearer that for advanced reactor designs that operate in the creep regime such as those designed in accordance and permitted by ASME Section III Division 5, that appropriate justification for flaw evaluation acceptance criteria shall be provided by the applicant.

(5) DG-1383 REFERENCE: v

Basis for Regulatory Guidance Position 4 (page 8 5th paragraph)

(5) BACKGROUND:

Similar to the background and recommendation 3 noted above, the use of the term refueling outage may not specifically have a universally understood definition for some advanced designs, that may not have a discrete refueling outage or cycle. Additionally, this Regulatory Guidance Position states that: *Licenseesshould submit the notification prior to the next refueling outage or within 3 years, whichever is less.”* This 3-year periodicity appears to be contradictory to the Regulatory Guidance Position offered in Regulatory Guidance Position 1 (page 7 - third bullet) which reflects a five year or less periodicity.

(5) RECOMMENDATION:

Provide a clarification to the presently used term *refueling outage* to be more inclusive to evolutions such as maintenance outages or a description of a time frame for the desired submittal of an OAR-1 Form, exclusive of using the terminology refueling outage, and (2) clarify what appears to be discrepancy in the desired periodicity reflected in the two noted Regulatory Guidance Positions.

(6) DG-1383 REFERENCE:

Basis for Regulatory Guidance Position 5 - Qualification of NDE personnel.

(6) BACKGROUND:

This Regulatory Basis is understandable since the staff has not to-date endorsed neither Code Case N-788-1, "Third Party NDE Certification Organizations Section XI, Division 1" nor the "ASME Nondestructive Examination and Quality Control Central Qualification and Certification Program," Standard, ANDE-1 2015. There are on-going discussions with staff to address their identified issues with the adoption of Code Case N-788-1.

The DG states, "The American National Standards Institute/American Society for Nondestructive Testing (ANSI/ASNT)-CP-189, "Standard for Qualification and Certification of Nondestructive Testing Personnel" (Ref. 13), may be used as provided in ASME Code, Section XI, Division 1." However, the DG does not address any of the many conditions associated with CP-189 included in 10 CFR 50.55a, ((b)(2)xviii) nor those addressed in Section XI, Division 1, IWA-2300. Additionally, it does not address the conditions related to Division 1, Mandatory Appendices VI, VII or VIII.

I am personally very perplexed that the NRC continues to impose the CP-189 Standard since there are many NRC reports, letters, etc., as well as, industry reports, that document the weaknesses of this program (Attachment 2, linked files). For example,(Attachment 2a), Ltr. Dated Jan. 12, 2010, to Kevin Ennis, Staff Secretary Board on Nuclear Codes and Standards The American Society of Mechanical Engineers, from Michele G. Evans, Director, Division of Component Integrity, Office of Nuclear Reactor Regulation, states, *"The Nuclear Regulatory Commission's (NRC's) Office of Nuclear Reactor Regulation (NRR) has become aware of an issue regarding activities associated with the American Society of Mechanical Engineers (ASME) and the American Society for Nondestructive Testing (ASNT) organizations. The issue regards questionable qualifications of Nondestructive Examination (NDE) personnel which relate to the ASME Boiler and Pressure Vessel Code (Code). Title 10 of the Code of Federal Regulations (10CFR) 50.55a(b) incorporates by reference the ASME Code, Sections III and XI. The Code addresses qualifications of NDE personnel; thus, these requirements become regulatory requirements with implications across the entire nuclear fleet. The NRR staff performed a review of the issue and believes that it would be better addressed by ASME. We are therefore forwarding this issue to your society for review and whatever action you deem appropriate. The NRC will follow ASME activities through our interactions on the Code committees."*

ASME responded on December 20, 2010,(Attachment 2b), stating, *"The American Society of Mechanical Engineers (ASME) appreciates the subject letter referenced above from the Nuclear Regulatory Commission's (NRC's) Office of Nuclear Reactor Regulation (NRR) bringing to our attention your concerns and issues regarding NDE personnel certification and probability of detection (POD). Please see the Attachment to this letter for specific responses to each of your questions under the issues addressed in your letter.*

ASME Codes and Standards are products of industry subject matter experts working together to reach consensus in establishment of requirements leading to the safe and reliable operation of pressure retaining components. In keeping with this tradition and successful process of standards development, the information you have provided will be submitted to the appropriate ASME Code Committee for consideration and action."

The nuclear industry in cooperation with ASME has introduced an initiative to develop a third-party NDE personnel certification program with the objective to consistently develop the technical workforce to meet current and future demands for NDE methods (UT, RT, MT, PT and VT) through a standardized centrally controlled process.

The ANDE-1 Standard, a performance-based third-party standard, has been developed as a direct response to these NRC concerns. Staff has been directly involved with the standard development and supported the implementation by contributions of funding and materials to develop UT practical examination samples.

In addition to ASME's support of the ANDE-1 standard, they initiated the Personnel Certification (PC), LLC, to serve as the Certifying Body to implement the standard. In this process they developed in cooperation of the Nuclear Specific Industry Sector (NSIS): a NQA-1 Quality Program; recognized a Designated Test Organization (DTO), Chattanooga State Community College; reviewed and audited a flaw sample supplier and a worldwide written examination administration organization; established an comprehensive computer-based system to receive candidate applications, review qualification requirements, and issue certificates; monitor requirements for re-certification; etc. Unfortunately, in January 2021, ASME had to discontinue the PC, LLC., based on the economic impact of COVID-19. However, as a part of their core business, ASME continues to support the ANDE-1 standard activities and published the ANDE-1 2020 update early in 2021. The NSIS Committee is actively continuing the development of the program and will make an announcement in the near future of the Certifying Body for full implementation. Included herein is the White Paper-Performance-Based NDE Personnel Certification dated July 30, 2015, (Attachment 2c) developed by the Section XI, Division 2, Working Group MANDE, to define and provide a technical basis for performance-based approach versus the deterministic approaches of ASNT-TC-1A, ANSI/ASNT CP-189, ISO-9712 and other similar schemes.

As addressed earlier, the NRC endorsement of the ANSI/ASNT CP-189 (2006) Standard as referenced in the current 50.55a is perplexing. The current rulemaking imposes four conditions on its use (b)(2)xviii. Additionally, ASME Section XI, IWA-2300 (2017) edition, imposes many more "amendments" or "band-aids" for implementation. Some examples include: (1) IWA-2310(b) that allows the ASNT ACCP program. It should be noted, that per ASME Interpretation XI-1-10-34, Issued 8/25/2011, (Attachment 2d) does NOT require the Owner or supplier to audit the neither the ASNT ACCP nor Level III certification program. ASNT's QA program is based on ISO 17024 not NQA-1 or Appendix B and only applies to the ASNT Level III certificate. Even the industry managed and funded Appendix VIII, Performance Demonstration Initiative (PDI) administered by EPRI is audited. (2) CP-189 requires in paragraph 2.2.1 NDT Level III requires certification as a ASNT NDT Level III or ACCP Professional Level III. Section XI, IWA-2314, states the ASNT Level III certificate is not required. This eliminates the third-party certification intended by CP-189 and allows employer certification of Level IIIs instead. (3) Training, qualification, and certification of visual and ultrasonic examination personnel shall also comply with the requirements of Mandatory Appendix VI and VII, respectively. However, unlike the ANDE-1 Standard, these are not performance-based and as addressed in the NRC's most recent report ML20079E343 (Attachment 2e), dated March 2020 and the EPRI Report EPRI_3002010462_(Attachment 2f)Tech Basis for App VII Code Presentation, dated October 29, 2021, the number and type of experience hours are inadequate for the required proficiency. (4) Additionally, as stated in the report ML20079E343, page 6, "These factors have led to a situation where trainees for Level II qualification gain some experience time on tasks such as making flow-accelerated corrosion (FAC) thickness measurements, but less experience performing other elements of a UT weld examination. Additionally, to increase the likelihood of successful PDI test performance, candidates are sometimes sent to EPRI to spend a week examining practice specimens prior to testing. Neither approach is optimal; making FAC measurements is a different task than UT weld examination, and concentrated practice before a test has been shown to be an ineffective strategy for long term knowledge maintenance."

(5) The NRC has removed the requirement that IGSCC Detection, Sizing and Weld Overlay examiners are required to attend specific training for each of these as was initially required per the IE Bulletin 82-03. (4) The NRC has removed the requirement that IGSCC Detection, Sizing and Overlay qualified via the EPRI PDI Program no longer have to re-qualify by examination every three-years. Data has shown that only ~50% of the personnel are successful on their first examination attempt. Additionally, no requalification examinations are required for other PDI qualifications.

The ANDE-1 Standard development is based on the tried and proved systematic approach to training (SAT) that was adopted by INPO to address issues after the Three Mile Island incident. It has been very successful in plant operator, maintenance worker, engineering staff, etc., It incorporates development of job-task analysis (JTA), body of knowledge (BoK), and experience documented via qualification cards. Further, instead on ill-defined hours of training and experience, it details via the qualification card exactly the knowledge and skills that have to be demonstrated prior to certification. As specified in the standard in addition to these requirements, Mandatory Appendix I, Nuclear NDE and QC Specific Industry Sector Requirements, comprise the detailed program requirements.

Initially, the US Nuclear utility industry voted to support the development of ANDE. This included providing some participation of development and monetary support. Support has dwindled as some have opted to preserve the employer self-certification CP-189 in lieu of implemention of performance-base third-party qualification and certification. There are however many other users, both domestic and international, of Section XI, that are actively participating in the implementation of the ANDE-1 NSIS.

If the USNRC continues with the endorsement of ANSI/ASNT CP-189, it is recommended that staff clarify in detail the conditions to be applied. Suggestions for these are detailed below in the RECOMMNEDATIONS section for ANSI/ASNT CP-189.

Performance demonstration of any MANDE that may be selected for an SSC under a RIM program is an essential and imperative input and quantitatively factors into the establishment a Reliability Target assigned to an SSC for such considerations as Probability of Detection (POD) criteria.

It is believed that while the USNRC has not formally endorsed to date the use of either Code Case N-788-1 or ANDE-1 2015, the reservation by the staff is understood to be based on the fact that ANDE-1 standard describes a qualification process of NDE personnel. If that is in fact the existing reservation, then the following recommendation is provided in the ANDE-1 RECOMMENDATIONS section.

(6) RECOMMENDATIONS:

ANSI/ASNT CP-189:
1a. Specify the conditions on Section XI, Division 1, addressed in 10 CFR 50.55a, (b)(2)(xv) and (b)(2).
1b. Reinstate the IE Bulletin 82-03 requirement that training for IGSCC detection, sizing and overlay weld repair examinations.
1c. Reinstate the requirement for 3-year recertification for PDI IGSCC detection, sizing and overly weld repair examinations.
1d. Require 3-year recertification for all Mandatory Appendix VIII PDI qualification examinations.
2. Condition the listed Section XI, Division 1, IWA-2300 amended requirements:
2a. **IWA-2314 Certification and Recertification** **"**except that the ASNT Level III certificate is not required." It is recommended that the third-party ASNT Level III certificate, ACCP certificate or another recognized third-party qualification with a NQA-1 or Appendix B QA program, e.g., EPRI NDE Center be required. Additionally, contrary to the ASME Interpretation XI-1-10-34 an audit of the ASNT Level III or ACCP certificate program would be subject to audit by the licensee.
2b. **IWA-2380 NDE INSTRUCTOR** **"**In lieu of the requirements of CP-189, "It is recommended the CP-189, third-party requirements are maintained as compared to those amended by IWA-2380.
2c. Additionally, the DG does not address the conditions related to Division 1, Mandatory Appendices VI, VII or VIII.

v

ANDE-1:
Provide guidance regarding the type of information that would be expected to be provided by a Licensee applicant that would allow for consideration of approval by the USNRC to use ANDE-1 for NDE personnel qualification under Division 2. This would assure that factors such as POD for any MANDE methods selected are established with consistency. The use of other already endorsed NDE personnal qualification criteria, such as CP-189 do not afford this essential criteria (i.e., POD) and potentially undermines a cornerstone to the development of a sound RIM Program.

(7) DG-1383 REFERENCE:

C. STAFF REGULATORY GUIDANCE 1. (PAGE 14 – 2ND BULLET AND 7TH BULLET)

(7) BACKGROUND:

It is understood that the USNRC would seek to have **summaries** of:
(a) The bases for the scope of the program and
(b) RIM strategies selected to achieve the reliability targets, as denoted in this portion of the Staff Regulatory Guidance at the time a Licensee/applicant makes an initial filing. However, both of these provisions may not be fully developed at the time of initial application and would not be fully vetted to be able to provide a detailed listing of all bases or strategies that may apply to each SSC selected to be within a final RIM program.

(7) RECOMMENDATION:

It is suggested that a clarification be provided to both of these items so as to better define what contents are expected to be provided by a Licensee/applicant at the time of license application.

(8) DG-1383 REFERENCE:

C. STAFF REGULATORY GUIDANCE 1. (PAGE 14 – 12^{TH} BULLET)

(8) BACKGROUND:

As previously outlined in comment 3 above, the FORM NIS-2 is merely an attestation for the completion of a Repair and Replacement Activity (RRA) signed by the Licensee and an Authorized Nuclear Inservice Inspector.

There is, however, no insightful or technical information contained on the FORM NIS-2 itself because the 2019 Edition of ASME XI Division 1 and Division 2 represents the incorporation of ASME Code Case N-532, where RRA that were required because an SSC had failed to an established ASME XI acceptance criteria would be documented on the OAR-1 Form, but the FORM NIS-2 became abbreviated.

(8) RECOMMENDATION: v

Consider deleting the requirement to submit NIS-2 Forms as suggested, since the information that is likely of interest to the USNRC will already be contained in the OAR-1 Form.

(9) DG-1383 REFERENCE:

C. STAFF REGULATORY GUIDANCE 5. *USE OF ANDE-1-2015 IS NOT ENDORSED BY THE NRC STAFF* (PAGE 15)

(9) BACKGROUND:

As cited in comment 6 above, performance demonstration of any MANDE that may be selected for an SSC under a RIM program is an essential and imperative input and quantitatively factors into the establishment a Reliability Target assigned to an SSC by using considerations such as Probability of Detection (POD) criteria. The use of ANDE-1 requires such performance demonstrations that are not a mandate of ANSI/ASNT CP-189.

(9) RECOMMENDATION:

Consider providing guidance regarding the type of information that would be expected to be provided by a Licensee applicant that would allow for consideration of approval by the USNRC to use ANDE-1 for NDE personnel qualification under Division 2 as an alternative to advocating the use of ANSI/ASNT CP-189.

Attachment 2

Attachment 2a – NRC LTR to ASME _1-12-20

Attachment 2b – ASME Response Ltr to NRC _ 12-20-2010

Attachment 2c – White Paper_Performance Based NDE Personnel Certification_7-30-2015

Attachment 2d – ASNT Audit Interpretation 11-262_8-25-2011

Attachment 2e-PNNL Report_ML20079E343_March 2020

Attachment 2f - EPRI_3002010462_Human Factors in NDE__ A Literature Review and Field Observations, October 2021

UNITED STATES
NUCLEAR REGULATORY COMMISSION
WASHINGTON, D.C. 20555-0001

Attachment 2a

January 12, 2010

Kevin Ennis, Staff Secretary
Board on Nuclear Codes and Standards
The American Society of Mechanical Engineers
Three Park Avenue
New York, NY 10016

SUBJECT: ISSUES RELATING TO THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS BOILER AND PRESSURE VESSEL CODE

Dear Mr. Ennis:

The Nuclear Regulatory Commission's (NRC's) Office of Nuclear Reactor Regulation (NRR) has become aware of an issue regarding activities associated with the American Society of Mechanical Engineers (ASME) and the American Society for Nondestructive Testing (ASNT) organizations. The issue regards questionable qualifications of Nondestructive Examination (NDE) personnel which relate to the ASME Boiler and Pressure Vessel Code (Code). Title 10 of the Code of Federal Regulations (10CFR) 50.55a(b) incorporates by reference the ASME Code, Sections III and XI. The Code addresses qualifications of NDE personnel; thus, these requirements become regulatory requirements with implications across the entire nuclear fleet. The NRR staff performed a review of the issue and believes that it would be better addressed by ASME. We are therefore forwarding this issue to your society for review and whatever action you deem appropriate. The NRC will follow ASME activities through our interactions on the Code committees.

Sincerely,

(/RA by M. G. Evans)

Michele G. Evans, Director
Division of Component Integrity
Office of Nuclear Reactor Regulation

Enclosure: As stated

January 12, 2010

Kevin Ennis, Staff Secretary
Board on Nuclear Codes and Standards
The American Society of Mechanical Engineers
Three Park Avenue
New York, NY 10016

SUBJECT: ISSUES RELATING TO THE AMERICAN SOCIETY OF MECHANICAL ENGINEERS BOILER AND PRESSURE VESSEL CODE

Dear Mr. Ennis:

The Nuclear Regulatory Commission's (NRC's) Office of Nuclear Reactor Regulation (NRR) has become aware of an issue regarding activities associated with the American Society of Mechanical Engineers (ASME) and the American Society for Nondestructive Testing (ASNT) organizations. The issue regards questionable qualifications of Nondestructive Examination (NDE) personnel which relate to the ASME Boiler and Pressure Vessel Code (Code). Title 10 of the Code of Federal Regulations (10CFR) 50.55a(b) incorporates by reference the ASME Code, Sections III and XI. The Code addresses qualifications of NDE personnel; thus, these requirements become regulatory requirements with implications across the entire nuclear fleet. The NRR staff performed a review of the issue and believes that it would be better addressed by ASME. We are therefore forwarding this issue to your society for review and whatever action you deem appropriate. The NRC will follow ASME activities through our interactions on the Code committees.

Sincerely,

(/RA by M. G. Evans)

Michele G. Evans, Director
Division of Component Integrity
Office of Nuclear Reactor Regulation

Enclosure: As stated

DISTRIBUTION
OAC Files
MCase

ADAMS ACC NO: ML100140091 Publicly Available/Non-Sensitive

OFC	AT:DE:NRR	DCI:NRR	DCI:NRR	DCI:NRR	OAC:NRR
NAME	Malave/Petrosino	CNove	TLupold	MEvans	GCwalina
DATE	01/07/10	01/07/10	01/08/10	01/11/10	01/12/10

OFFICIAL RECORD COPY

REQUEST FOR REVIEW

Concern:

There may be a generic issue with the effectiveness of nondestructive examination (NDE) inspections at all operating nuclear power plant stations in the United States. Specifically, the personnel who perform nondestructive examinations in the nuclear power industry may not be properly trained, and do not have the necessary education in mathematics, metallurgy and materials technology to perform their jobs adequately. As a result, the probability of detection (POD) of flaws may be as low as 50% using general methods. Of particular relevance to ASME, the American Society for Nondestructive Testing (ASNT) nondestructive testing (NDT) personnel certifications may not be typically acceptable per code expectations. Scores on some ASNT Level III examinations may have been lower than the Code required 70% for individual components and a composite score of 80% for all portions combined; however, due to the methods ASNT used to grade exams including the use of psychometrics, ASNT would have considered the scores to be passing. Thus, personnel with potentially inadequate test results could meet the qualification requirements of the ASME Code.

Issue 1:

For in-service inspection, personnel qualification requirements are found in the ASME Code, Section XI, IWA-2300 "Qualification of Nondestructive Examination Personnel." For new construction, personnel qualification requirements are found in Section III, NX-5520, "Personnel Qualification, Certification and Verification." Conditions, such as those implemented by 10 CFR 50.55a(b)(2)(xviii) "Certification of NDE Personnel" and 10 CFR 50.55a(b)(2)(xiv) "Appendix VIII Personnel Qualification," provide additional requirements. ASME Code, Section XI, IWA-2310(b) "Qualifications of Nondestructive Examination Personnel," endorses the use of the ASNT Central Certification program (ACCP) as an alternative to the personnel qualification program based on ANSI/ASNT CP-189. ASME Code, Section III, NB-5521(a) requires the use of ASNT SNT-TC-1A for personnel qualifications.

Relative to these Code requirements for qualifying personnel the NRC requests ASME to consider the following questions:

A. Are there deficiencies with the quality of qualification and certification of NDE personnel given that ASNT certifications (ANSI/ASNT CP-189, ACCP and SNT-TC-1A) may potentially fall short, specifically in regards to the methods implemented by ASNT to conduct testing, grade examinations and report results for interpretation by the Code?

B. How does ASME ensure that ASNT's processes meet the Code requirements including adjustments of test scores by use of psychometrics or other means?

C. Has ASME audited ASNT's certification process to ensure compliance?

Issue 2:

NRC considers inservice inspection, and thus POD for the inservice inspection methods, to be important components of the defense-in-depth approach to nuclear power plant safety. As mentioned previously, POD may be as low as 50% using general methods. Please address how the POD is factored into the ASME Code's determination of Code criteria. For instance, since the joint efficiency (Section III, Paragraph UW-12) is based, in part, upon the degree of examination of the joint, is POD for radiographic examination a factor in determining the Code criteria?

If you have any questions on this request for review, please contact Carol Nove at (800) 368-5642.

Attachment 2b

ASME
SETTING THE STANDARD

STANDARDS & CERTIFICATION

Three Park Avenue
New York, NY
10016-5990 U.S.A.

tel 1.212.591.8500
fax 1.212.591.8501
www.asme.org

December 20, 2010

Ms. Michele G. Evans, Director
Division of Component Integrity
Office of Nuclear Reactor Regulation
Mail Stop 9H6 One White Flint North
11555 Rockville Pike, Rockville, MD 20852
Michele.Evans@nrc.gov / (301) 415-2795

Subject: Response to letter from NRC to ASME on NDE Issues Relating to the ASME Boiler and Pressure Vessel Code

Reference: Letter from Michelle G. Evans, NRC, to Kevin Ennis, ASME, dated January 12, 2010, Subject: Issues Relating to the ASME Boiler and Pressure Vessel Code

Dear Ms. Evans:

The American Society of Mechanical Engineers (ASME) appreciates the subject letter referenced above from the Nuclear Regulatory Commission's (NRC's) Office of Nuclear Reactor Regulation (NRR) bringing to our attention your concerns and issues regarding NDE personnel certification and probability of detection (POD). Please see the attachment to this letter for specific responses to each of your questions under the issues addressed in your letter.

ASME Codes and Standards are products of industry subject matter experts working together to reach consensus in establishment of requirements leading to the safe and reliable operation of pressure retaining components. In keeping with this tradition and successful process of standards development, the information you have provided will be submitted to the appropriate ASME Code Committee for consideration and action.

The nuclear industry in cooperation with ASME has introduced an initiative to develop a third party NDE personnel certification program with the objective to consistently develop the technical workforce to meet current and future demands for NDE methods (UT, RT, MT, PT and VT) through a standardized centrally controlled process.

Again we thank you for bringing these issues to our attention

If you have any questions regarding the contents of this letter please contact me or Mr. Kevin Ennis, ASME Director, Nuclear Codes and Standards by telephone (212) 591-7075 or by e-mail *ennisk@asme.org*.

Very Truly Yours,

Bryan A. Erler

Bryan A. Erler, PE
Vice President
ASME Nuclear Codes and Standards
erlerltd@aol.com / (773) 248-6849

cc: Members, ASME Board on Nuclear Codes and Standards
Members, ASME BPV Committee on Nuclear Inservice Inspection (XI)

ATTACHMENT

Response to NRC Questions Stated Under Issue 1

Question A: Are there deficiencies with the quality of qualification and certification of NDE personnel given that ASNT certifications (ANSI/ASNT CP-189, ACCP and SNT-TC-1A) may potentially fall short, specifically in regards to the methods implemented by ASNT to conduct testing, grade examinations and report results for interpretation by the Code?

Response:
All employers must certify their NDE personnel as meeting minimum Code requirements. However, employers may use outside agencies to administer some or all of the examinations. It is the employer's responsibility to audit third party providers of examination services to assure compliance to Code requirements.

Question B: How does ASME ensure that ASNT's processes meet the Code requirements including adjustments of test scores by use of psychometrics or other means?

Response:
The ASME BPV Code does not contain requirements related to adjustments of scores for examinations by use of psychometrics or other means. Outside agencies are allowed to issue pass/fail results in accordance with IWA-2323(f) and VII-4350(a) and employers shall assign 80% for passing results.

Question C: Has ASME audited ASNT's certification process to ensure compliance?

Response:
As stated above, audits are the responsibility of the employer.

Response to NRC Question Stated Under Issue 2

Question: NRC considers inservice inspection, and thus POD for the inservice inspection methods, to be important components of the defense-in-depth approach to nuclear power plant safety. As mentioned previously, POD may be as low as 50% using general methods. Please address how the POD is factored into the ASME Code's determination of Code criteria. For instance, since the joint efficiency (Section III, Paragraph UW-12) is based, in part, upon the degree of examination of the joint, is POD for radiographic examination a factor in determining the Code criteria?

Response: No, POD is not applied to the ASME radiographic examination criteria. This is because the weld must meet Section IX and NX-4000 which have been established as satisfactory. The history of safe and successful operation of pressure retaining components designed and constructed to ASME Code requirements is testimony to its sound approach to design, and its requirements for NDE.

There appears to be some confusion on this issue as indicated by the example presented, as UW-12 is a paragraph in Section VIII, not Section III.

Defining Performance-Based NDE Personnel Qualification

Background & Introduction

This establishes an alternative approach for nondestructive examination (NDE) personnel qualification for the boiler and pressure vessel industry.

Based on quantitative NDE reliability studies conducted to-date [1][2], the current deterministic approach to NDE personnel qualification based on such schemes as ASNT SNT-TC-1A, ANSI/ASNT CP-189, EN-473, ISO-9712 and other similar approaches are not as effective as desired. The goal of this document is to present an alternative approach to the deterministic NDE personnel qualification schemes.

Personnel are commonly qualified for technical tasks by completing a specified amount of classroom training hours and time on the job (on-the-job training). However, training and job experience does not necessarily lead to job proficiency. For example, research involving 32 candidates (two groups) for qualification in the Electric Power Research Institute (EPRI) Nondestructive Evaluation (NDE) Center Performance Demonstration Initiative (PDI) program revealed no positive relationship between the number of years of NDE, ultrasonic, or piping examination experience and performance on quantitative ultrasonic examination demonstration tests.[3] Results such as these are not puzzling if one recognizes that unstructured training and job experience may or may not include the key factors that are essential for learning: measurement and feedback of performance results. It is difficult to measure performance and provide feedback on-the-job since actual flaws are rarely encountered or if they are, feedback is rarely provided. Timely, accurate measurement and feedback may be difficult even in laboratory-type blind demonstration tests but if demonstration tests of similar difficulty and conditions are provided in a structured systematic approach to training (SAT) then the skills can be more effectively learned by the candidates. This performance-based training will provide the desired results of improving the personnel NDE skills.

The report, *Nondestructive Evaluation: A Review of NDE Performance Demonstrations – NDE Round Robin Report.* EPRI, Palo Alto, CA: 2008. 1016969,[3] (Note: this report is publicly available at (www.epri.com/search/Pages/results.aspx?k=1016969) published in June 2008, provides a summary of various performance demonstrations conducted worldwide to determine the ability of certified nondestructive evaluation (NDE) personnel to pass performance-based practical tests and demonstrations. These tests were conducted throughout the world for a variety of industries. In general, the pass rates for these other blind tests have been comparable to the U.S. nuclear experience of about 50%. As detailed in the abstract, “This report describes a study performed by the Electric Power Research Institute (EPRI) Nondestructive Evaluation (NDE) Program to assist the Tennessee Valley Authority in providing a basis for performance-based NDE personnel certification. History has shown that re-qualification examinations for intergranular stress corrosion cracking (IGSCC) ultrasonic (UT) examiners have pass rates averaging 57%. The U.S. Nuclear Regulatory Commission (NRC) staff has questioned whether examiners have maintained their proficiency over the three-year re-qualification period and may recommend more frequent and more stringent testing requirements. An alternative to stringent re-qualification requirements is to provide annual practice or training as part of a performance-based qualification approach. This report will summarize the past performance of experienced candidates that have been certified to different written practices and programs and indicate their capabilities as measured by undertaking hands-on practical qualification testing. Each chapter describes a study and includes an excerpt from the study's documentation." Specifically, the EPRI Report 1016969 includes information from the following:

- *Nondestructive Evaluation (NDE) Capabilities Data Book;* this is for engineering analyses in the form of a condensed reference to previously demonstrated NDE capabilities for the aircraft industry.

- *Reliability of Nondestructive Inspection (NDI) of Aircraft Components;* this report documents and concludes an initial task to assess the reliability of nondestructive inspection (NDI) of aircraft engine components in Air Force engine overhaul facilities.

- *API Qualification of Ultrasonic Examiners Certification Program*, American Petroleum Institute. This study is intended to provide an outline of the API Qualification of Ultrasonic Examiners Certification Program. The American Petroleum Institute's performance based certification examinations for ultrasonic examination are performed in support of petro-chemical plant integrity. As part of their Individual Certification Program (ICP) examinations are given in four techniques 1) manual conventional ultrasonic flaw detection and characterization, 2) ultrasonic flaw height sizing, 3) tank floor corrosion mapping and 4) phased array ultrasonics. To be allowed to take examinations a minimum of an SNT-TC-1A level II (employer-based qualification) is required. Approximately 300 technicians have completed certification to some portion of the program as mandated by API member organizations and API standards. Examinations are administered by API using a third-party examination organization. While there are no formal and published statistics for successful first attempt completion, it is reputed to be no greater than 50%.
- Program for the Assessment of NDT in Industry (PANI) Seminar Report by John Thompson; British Institute of Non-Destructive Testing (BINDT)/PCN, taking into account the Personnel Certification in Non-Destructive Testing (PCN) recertification experience, where the majority of candidates undergo some form of refresher training prior to examination, and even so, failure rates are about 30% to 40%.

- Jack Spanner, Jr. “White Paper Report 1: What is the Real Quality of ASNT Certified Personnel?” 2002 ASNT Fall Conference Panel; In general, the pass rate for IGSCC detection on the first attempt was between 19% and 34% and this increased to 27% and 70% after multiple attempts. This data is based on 27 classes with 182 candidates held from September 1985 until July 1987 [1]. Only 18% of the candidates passed the combined General, Specific and Practical tests on the first attempt. The IGSCC qualified personnel are still required to pass the IGSCC demonstration test without the written tests every three years and their pass rate is approximately 50% on the first attempt. This increases to approximately 75% after multiple attempts. All of the participants in these two programs are required to be fully certified to ASNT Level II or Level III in the UT method.

The results of the EPRI review state,

> *“In general the pass rates for these other blind tests have been comparable to the nuclear domestic experience. That is, the pass rates obtained by experienced certified personnel is very rarely above 50% on their first attempt at passing a demonstration test that simulates the examinations that they conduct in the field. Some of the test mockups with flaws were even removed from service so there can be no question about their applicability to field inspections.”*

> *“These pass rates have a large impact on the costs to maintain the qualification of NDE personnel. Each attempt by a candidate to pass a demonstration test incurs “tuition”, labor, and usually per-diem costs. Candidates that don’t pass the qualification tests may not be hired for nuclear work so the individual loses income and the employer is not compensated for preparation training that was provided. There is room for improvement in the area of effectively qualifying NDE personnel. Adopting performance-based concepts to training, experience and testing of NDE personnel in the future should improve the reliability of examinations in the field.”*

Similar results were documented in the KARTA report, Three Decades On NDI Reliability Assessment, Karta -3510-99-01.

> “*This report reviews three decades (1970 – 1999) of engineering and research efforts to quantify capability and reliability of non-destructive inspections in aerospace and other industries… The task was performed during July-October 1999 and covers nearly 150 reports and manuscripts from over 100 authors. Experts on the subject reviewed the report during December 1999 – March 2000.”*

The KARTA report, “Summary of Observations” addresses programs, including one of the first quantitate studies, “Have Cracks Will Travel.” One of these program observations was, “Programs do not appear to evaluate what best could have been achieved; in some sense, the ideal capability measurement did not receive proper attention. Recent programs do not show clear evidence of using experience from earlier programs." Additionally, the summary of observations addresses, specimens, where it was noted, "Specimen number and design, fabrication, flaw induction, flaw characterization, and maintenance hold the key to success of the program. Generally, the specimens have been artificially fabricated specifically for the purpose of assessment. Some programs had a combination of synthetic and actual flawed parts." Relative to Human Factors, the report stated, "Many investigations believe that if NDI engineering is not well defined, the human should not be blamed for poor performance. There continues to be a lack of good human factors investigation and mitigation programs."

The report also states in the summary, "Most of the studies identified controllable as well as non-controllable factors that clearly effect NDI reliability. These include the material, process, equipment, procedure, and more importantly the human. Many investigations attributed the variations observed in detection capability to the human element. Human factors have become a subject of great interest and deserve an in-depth investigation." This further supports the need for an improved approach for the qualification of NDE personnel.

A number of professions use a performance-based approach to assure that personnel performing critical task are effective. These include commercial and military aircraft pilots, nuclear plant maintenance workers and plant operators, information technology (software and system) professionals, teachers, fire fighters, forestry loggers, welders, utility linemen, etc. The use and acceptance of a performance-based approach is rapidly increasing as a process to measure critical job task competency. The concept of performance-based qualification is very straight forward, i.e., personnel demonstrate they are capable of performing tasks to produce outputs to meet defined requirements by doing the tasks under a set of conditions where this capability can be quantitatively measured. Although the concept is straight forward, many complex challenges exist that must be identified, developed and implemented to achieve an effective performance-based approach.

The current NDE qualification process includes some of the essential elements of a performance-based qualification, e.g., the hands-on practical examination that is evaluated by a NDE Level III individual. However, based on the quantitative studies conducted, the current approach has not been successful. The current qualification process requires prescriptive amounts of classroom training, on-the-job experience as well as successful completion of written examinations to assess knowledge and a hands-on practical examination to demonstrate skill. Why then are these time tested elements not providing the desired performance results and what can be done to improve the results?

Organizations in the United States (US) and many other countries are using the American Society of Nondestructive Testing's (ASNT) Recommended Practice No. SNT-TC-1A, as a guideline for employer-based training, qualification and certification. Employer-based certification using a guideline results in very large variations in the quality of the training, experience and qualification examinations used as the basis for certifications. A number of the American Society for Mechanical Engineers (ASME) Boiler and Pressure Vessel Code (ASME B&PVC) sections still adopt this document as the required basis for certification of NDE personnel. Some ASME B&PVC sections, e.g., Section XI, have referenced the ANSI/ASNT CP-189 Standard for employer based certification along with including additional requirements for NDE personnel certification. Results of some quantitative studies, since implementation of these requirements, i.e., SNT-TC-1A and ANSI/ASNT CP-189, still document low performance. There are still significant variations in the employer-based training, experience and qualification examinations administered. The differences of these approaches with a performance-based approach to training, experience and examination issues will be addressed below.

Performance-based Training

The guidelines, Code(s) and regulations have prescribed a certain minimum number of classroom training hours for individuals to be eligible to become qualified. The classroom training hours required to address a specific NDE method and technique(s) theoretical topics needed to gain the knowledge to conduct the NDEs are specified. However, there is a significant amount of variance in the quality and content of the materials actually presented and significant variation in the knowledge and skills of the individuals conducting the training. The Institute of Nuclear Power Operations' (INPO) systematic approach to training (SAT) has demonstrated a more successful approach, e.g., INPO's Nuclear Maintenance Workers and Plant Operator's Programs. The SAT approach includes the elements of analysis, design, development, implementation and evaluation. Other commonly used terms for the SAT approach are Training System Development (TSD), Analysis, Design, Development, Implementation, and Evaluation **(**ADDIE) and Accomplishment-Based Curriculum Development (ABCD**).** Training developed with the SAT model depends upon systematic movement through all five phases at least once or more than once, if revision is necessary. The evaluation phase determines if training was successful, how successful it was, and if any problems are detected and how to correct the problems. Evaluation is the SAT phase that ties all other phases together through feedback. The outcome of one phase becomes input for the next. Feedback ensures that the transition of training through the phases stays on course.

These variations in the quality and content of most NDE training can be addressed by developing "standardized" student and instructor training course materials that are required to be used for the training. Such materials should be provided to NDE Instructors who would be trained to use them. Since individuals learn at different rates, it is not the most effective use of human resources to require a prescriptive minimum number of classroom training hours. Implementation of “standardized” training materials that may be delivered in the traditional classroom method, on-line, computer-based learning, or self-paced learning, can be much more effective and ensures the knowledge-based information is comprehensively addressed. SAT is supported with psychometrically validated written examinations to effectively assess the comprehension of the knowledge. Simply stated, psychometrics is the science and technology of mental measurement. An example that most have taken is an intelligence quotient (IQ) test. Psychometrics also applies to instruments (tests) that measure skills, i.e., performance examinations, hands-on practical, and assessment instruments, for example, the Minnesota Multiphasic Personality Inventory (MMPI), etc. Psychometrics provides a scientific process to move from a traditional prescriptive approach to a statistically based standard evaluation of personnel knowledge and skills.

Relevant Experience

As with training, the guidelines, Code(s) and regulations have prescribed a certain minimum number of experience hours (on-the-job training) for individuals to be eligible to become qualified. These requirements do not explicitly describe the specific tasks to be performed to obtain credit for the experience hours. As proposed for the performance-based training, the experience needs to be more clearly defined and standardized. For example, for a candidate to become fully certified as a Level II in the magnetic particle method, they should be required to examine an array of test items, e.g., product forms such as castings, forgings, plate, pipe, etc., and welds made using various welding processes, e.g., SMAW, GTAW, GMAW, etc., using all the various magnetic particle examination techniques. The current practice does not assure that a candidate with the minimum number of experience hours has had this type of experience. Of course, if “limited” qualification is desired, this too should be detailed as to exactly what this shall consist. A performance-based approach would describe the experience needed to acquire the necessary skills through a job-task-analysis for NDE methods and techniques.

What is a Performance-Based Qualification/Certification System?

A performance-based system verifies and documents that a person is capable of performing tasks to produce outputs to meet defined industry requirements. These defined requirements should be developed based on job-task-analysis by subject matter experts for each respective industry.

Compliance with the defined requirements is accomplished by having the candidate either performing the task while being observed and questioned by a qualified assessor and/or assessing the quality of the end product and perhaps interviewing stakeholders and/or others involved in the performance. The assessor uses a form that contains standards for the procedures that must be followed and the criteria for assessing acceptable performance. The more comprehensive way of conducting the assessment is to combine the approaches, i.e., use a qualified assessor with an objective based criteria and have the candidate perform the task on the product so a quantitative result can be measured. The completed test form becomes the documented record of qualification for the task.

The battery of knowledge and skills tests covers all the critical tasks the individual must be able to perform to the established level of competence in a specific role such as Control Room Operator, Sales Representative, or Six Sigma Black Belt practitioner. When an individual has qualified on all the tasks for an assigned role, they can be said to be “fully qualified” or “certified” for that role.

How does the performance-based approach differ from other approaches? The performance-based approach requires that candidates demonstrate that they can actually perform the work to a defined standard - this is what is meant by performance-based. Some other systems use a prescriptive approach and test for knowledge, or give tests about the work.

Demonstrated task performance is a more reliable way to know that candidates can successfully perform the work. Knowledge testing is simply not adequate. Some candidates who can perform the work with a high degree of proficiency may not be good test takers and may actually fail on work they have really mastered. Conversely, a good test taker can pass a knowledge test about the work without actually being able to perform. Additionally, good test takers may acquire the minimum skill with an element of luck to pass a performance test, if the performance test is not properly structured to minimize these elements.

Another approach requires specific training in order to be “certified.” This approach is very weak in terms of assurance that the individual can really perform specific tasks unless the performance of the tasks is actually observed (under job-realistic conditions), evaluated, and documented in the training program for each candidate. Following are some examples of other professions that have adopted a performance-based approach for certification:

The Red Hat Certified System Administrator – RHCSA, **“**The certification exam is performance-based, meaning that candidates must perform tasks on a live system, rather than answering questions about how one might perform those tasks.” [4]

Institute for Credentialing Excellence; *Setting Passing Standards for Performance-Based Certification & Licensure Examinations.* [5]

> *Certification is a process by which a nongovernmental agency validates, based upon predetermined standards, an individual nurse's qualification and knowledge for practice in a defined functional or clinical area of nursing.*
>
> *Certification validates your knowledge of nursing in your specialty area to hospitals, peers, patients and, most importantly, to yourself. Certification promotes continuing excellence in the nursing profession.*

Performance-based teacher certification in Georgia [6] is centered on:

> *criterion-referenced tests, and on-the-job assessment procedures for student teachers and beginning teachers. A state-funded contract was awarded to develop teacher certification tests, resulting in a 250-item pool and 15 criterion-referenced tests for 32 teaching fields. Cutting scores were derived from information supplied by Georgia experts who reviewed test items for minimum content knowledge/competence. Five instruments were developed and tested: Teaching Plans and Materials, Classroom Procedures, Interpersonal Skills, Professional Standards, and Pupil Perceptions. Master teachers were trained as data collectors. Performance based certification will be implemented in 1980, through the establishment of a statewide network of regional assessment centers. Full implementation of the program will be phased in: administration of appropriate teaching-field criterion referenced tests; assessment of student-teaching performance by college supervisors and classroom supervising teachers; and conversion of nonrenewable teaching certificates to renewable certificates through satisfactory on-the-job performance as measured by the Teacher Performance Assessment Instruments: Beginning Teacher Form. (MH)*

Performance competence is the ability to perform tasks to produce outputs to stakeholder requirements. [7]

According to a medical study entitled, The Effects of Practicing with a Virtual Ultrasound Trainer on Fast Window Identification, Acquisition, and Diagnosis **[8]** states,

> *Unfamiliarity with ultrasonography, the cost of training users on ultrasound-guided procedures, and the lack of training opportunities are limiting the use of this potentially beneficial technology. A cost-effective and widely applicable method for providing users with ultrasound-guided procedural training is needed. Purported simulator-based training advantages over traditional medical training include: (a) no risk to trainees or patients during practice attempts; (b) more cost-effective than current training methods; (c) provides multiple modes of sensory interaction to maximize learning; (d) user gains and maintains proficiency through*

unlimited repetition (avoids skill decay); (e) independent, self-directed learning; and (f) provides a method for performance tracking. In this study, we compared simulation-based practice of ultrasound scanning to classroom-based practice of ultrasound scanning on knowledge and performance measures. We focused on one type of procedure, the focused assessment with sonography for trauma (FAST) exam, as the context for the comparison.

> *The study reports, "Participants learned over instruction and practice on all knowledge measures, with percent gains of 121%, 133%, 474%, and 535% on basic FAST scanning procedures, anatomical interpretation of FAST windows, identification of FAST window quadrants, and diagnostic interpretation of FAST windows, respectively."*

According to another study entitled, *Review of Rifle Marksmanship Training Research*, [9]

> *Device-fire performance, or shooting performance on a rifle simulator, has been one of the strongest predictors of record-fire performance. The use of rifle simulators has received much attention because of the cost-savings potential for sustainment-level training and remediation.*

The report, *Technical Issues in Large-Scale Performance Assessment* [10] is another example that supports the cost benefits of a performance-based approach to learning and assessment. It states,

> *The issue of utility is evaluated in terms of the benefits or desired outcomes of the assessment relative to its costs (Cronbach & Gleser, 1965; Messick, 1989). Thus, although the cost of performance assessments in terms of time and resources is an important consideration, the choice among alternative assessment approaches should not be determined solely by cost or efficiency. Rather, such decisions should weigh both the costs and the benefits of the assessment, that is, its utility for the applied purpose.*

Organizations, such as Prometric's, develop and deliver performance-based qualifications for a large number of professions including those associated with ASME in non-pressure vessel related applications. As shown from their website,

> *"Prometric* [11] *is a test development and delivery provider to more than 400 organizations worldwide. Prometric's securely delivers an average of 10 million exams per year."*

Such organizations use psychometrically validated computer-based tests (CBT) as a way to make performance-based qualification competitive with traditional test delivery.

The ASME standard, ASME ANDE-1-2013, *ASME Nondestructive Examination and Quality Control Central Qualification and Certification Criteria,* is being developed to provide the performance-based requirements for use by industry. This white paper is intended to support development and implementation of this standard.

Summary

A systematic approach to training together with performance-based tests, psychometrically validated evaluation of knowledge and skills will improve a candidate's performance. A majority of traditional employer-based written examinations are not developed or validated psychometrically. The use of third-party psychometrically validated examinations would replace the current practice of employer developed and administered examinations. It improves upon traditional ASNT SNT-TC-1A, ASNT/ANSI CP-189, ISO-9712, etc., requirements by including more comprehensive hands-on practical examinations on a statistically valid set of samples containing flaws representative of those expected to be encounter in shop and field conditions. The sample sets will be designed for either a "general" NDE method, or "limited" technique(s) of a method, or for industry specific sector needs, as applicable.

Acronyms

ABCD – Accomplishment-Based Curriculum Development

ADDIE – Analysis, Design, Development, Implementation, and Evaluation

ANSI – American National Standards Institute

ASME – American Society for Mechanical Engineers

ASNT – American Society of Nondestructive Testing

EPRI – Electric Power Research Institute

IGSCC – Intergranular Stress Corrosion Cracking

ISO- International Standards Organization

NDE – Nondestructive Examination

NDE – Nondestructive Evaluation

NDI – Nondestructive Inspection

NDT – Nondestructive Testing

PDI – Performance Demonstration Initiative

SAT – Systematic Approach to Training

TSD – Training System Development

References

1. *Nondestructive Evaluation: A Review of NDE Performance Demonstrations–NDE Round Robin Report.* EPRI, Palo Alto, CA: 2008. 1016969.

2. Three Decades of NDI Reliability Assessment, KARTA -3510-99-01, May 2000, Ripudaman Singh

3. Stephens, Henry, NDE Reliability – Human Factors Basic Considerations,15th World Conference on NDT, Rome Italy, May, 2000, IDN-736

4. http://www.redhat.com/training/certifications/rhcsa/

5. http://www.credentialingexcellence.org/Store/tabid/432/pid/6/Setting-Passing-Standards-for-Performance-Based-Certification-Licensure-Examinations.aspx

6. http://www.eric.ed.gov/ERICWebPortal/search/detailmini.jsp?_nfpb=true&_&ERICExtSearch_SearchValue_0=ED179605&ERICExtSearch_SearchType_0=no&accno=ED179605

7. http://www.aacn.org/WD/Certifications/Content/generalinfo.content?menu=Certification&lastmenu=

8. Chung, G. K. W. K., Gyllenhammer, R. G., & Baker, E. L. (2011). The effects of practicing with a virtual ultrasound trainer on FAST window identification, acquisition, and diagnosis. (CRESST Report 787). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

9. Chung, G. K. W. K., Nagashima, S. O., Delacruz, G. C., Lee, J. J., Wainess, R., & Baker, E. L. (2011). *Review of rifle marksmanship training research.* (CRESST Report 783). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

10. Edited by Gary W. Phillips, Associate Commissioner, National Center for Education Statistics, Technical Issues in Large-Scale Performance Assessment, NCES 96-602, April 1996

11. http://www.prometric.com/reference/PBT.htm

12. "Career Development for Linemen Program Manual & Enrollment Information,"Copyright 1999 by Minnesota Rural Electric Association 11640 73rd Ave. N. Maple Grove, MN 55369

Interpretation Detail

Attachment 2d

Standard Designation:	BPV Section XI
Edition/Addenda:	
Para./Fig./Table No:	
Subject Description:	IWA-1400 and IWA-2300 (1989 Edition Through the 2010 Edition With the 2011 Addenda)
Date Issued:	08/25/2011
Record Number:	11-262
Interpretation Number:	XI-1-10-34
Question(s) and Reply(ies):	Question: Is it a requirement of IWA-1400 or IWA-2300 that, for the Owner or employer to accept ACCP or Level III certificates issued by the ASNT, that the Owner or employer audit the ASNT as a supplier? Reply: No.

Attachment 2e

Pacific Northwest
NATIONAL LABORATORY

PNNL-29761

Nondestructive Examination (NDE) Training and Qualifications: Implications of Research on Human Learning and Memory, Instruction and Expertise

March 2020

TF Sanquist

Prepared for the U.S. Nuclear Regulatory Commission
under a Related Services Agreement with the U.S. Department of Energy
CONTRACT DE-AC05-76RL01830

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

PNNL-29761

Nondestructive Examination (NDE) Training and Qualifications: Implications of Research on Human Learning and Memory, Instruction, and Expertise

March 2020

TF Sanquist

Prepared for
the U.S. Nuclear Regulatory Commission
under a Related Services Agreement with the U.S. Department of Energy
Contract DE-AC05-76RL01830
CA Nove, Contracting Officer Representative

Pacific Northwest National Laboratory
Richland, Washington 99354

Abstract

This report and the associated appendices present a review of the literature concerning human learning, memory, training, and instructional psychology as it pertains to issues of training and qualification in nondestructive examination. The literature has been synthesized into fifteen fundamental principles of learning, and guidance for application of each is discussed. Specific recommendations for laboratory practice and field experience time are made for ultrasonic testing, but also have broader applicability to many skill-based jobs. The principles of learning suggest that increased focus on detection and characterization of flaws with lab practice can enhance examiner skill, in conjunction with a sufficient amount of field experience to develop proficiency in operational settings.

Acronyms and Abbreviations

AED	automated external defibrillators
ANDE	ASME NDE
ARI	Army Research Institute
ASME	American Society of Mechanical Engineers
ASNT	American Society for Nondestructive Testing
BMT	behavior modeling training
BPV Code	Boiler and Pressure Vessel Inspection Code
CME	continuing medical education
CPR	cardiopulmonary resuscitation
CTE	career and technical education
DBT	demonstration-based training
EPRI	Electric Power Research Institute
FAA	Federal Aviation Administration
FAC	flow-accelerated corrosion
FIT	Feedback Intervention Theory
ICAP	Interactive, Constructive, Active, Passive
MOC	maintenance of certification
NDE	nondestructive evaluation/examination
NDT	nondestructive testing
NRC	U.S. Nuclear Regulatory Commission
OE	operating experience
OJT	on-the-job training
PDI	Performance Demonstration Initiative
PRS	Perceptual Representation System
SME	subject matter expert
SOI	Science of Instruction
SOL	Science of Learning
STEM	science, technology, engineering and mathematics
TAU	thinking, asking, and understanding
TER	transfer effectiveness ratio
UDA	User’s Decision Aid
UT	ultrasonic testing
VR	virtual reality

Contents

Figures

Tables

1.0 Introduction

This technical letter report presents a synopsis and integration of a literature review of scientific research in human learning and memory, training, and development of expertise. The review was undertaken to support development and evaluation of a technical basis for field experience hours to qualify ultrasonic testing (UT) examination personnel as Level II examiners.

The training and qualification requirements for examiners are specified in the American Society of Mechanical Engineers Boiler and Pressure Vessel Inspection Code (ASME BPV Code), Section XI, Appendix VII. These requirements are based on the American Society for Nondestructive Testing (ASNT) Recommended Practice No. SNT-TC-1A, “Personnel Qualification and Certification in Nondestructive Testing,” which was developed in the 1960s. To overcome variations in implementation of recommended practices across employers, ASNT developed CP-189 “ASNT Standard of Qualification and Certification of NDE Personnel,” which describes the minimum requirements of a certification program, based on national consensus. Experience hour requirements for Level II UT certification vary between SNT-TC-1A and CP-189; the former recommends 840 hours minimum with 1200 total in nondestructive evaluation (NDE), while the latter requires 630 minimum with 1200 hours total. The ASME BPV Code Appendix VII is based on CP-189, and until 2011 required 800 hours of work experience for UT certification. Code Case 784-1 reduced the experience requirements to 320 hours of laboratory practice and 80 hours of field experience for UT certification. Currently, 10 CFR 50.55a prohibits the use of the reduced hours because the effect of reducing operational experience is unknown. There is no guidance in Code or regulation for what constitutes field experience, and it has historically been open to wide interpretation.

Common practice among vendors is to ensure that their Level I personnel accrue sufficient field experience hours to meet the Appendix VII requirements. These requirements also include passing an examination using specimens in a blind testing environment. These tests are often administered by the Electric Power Research Institute (EPRI) under the Performance Demonstration Initiative (PDI); passing these tests results in an examiner being “PDI-qualified.” It is also common practice for vendors to send examiners to EPRI prior to testing so that they can practice UT scanning on specimens similar to those used for performance demonstration. The rationale of the Appendix VII requirements for training, field experience, and testing (performance demonstration) is that the combination of these elements will produce proficient NDE examiners.

Field experience is thought to facilitate the development of proficiency in the job knowledge and skills to competently perform examinations in the operational setting, and detect and characterize flaws in nuclear power plant components. More generally, effective performance of job skills depends on various types of memory and knowledge of the employee. Memory and knowledge for application on the job is established through diverse learning processes and training processes.

There are certain practical considerations associated with field experience that warrant consideration. The most significant consideration is the low rate of occurrence of flaws in operational settings. Detecting and characterizing flaws and indications is the most critical part of the examiner’s job; but in the field, it is not necessarily exercised in a way that will allow for skill development. Thus, potential tradeoffs between laboratory practice and field experience are being evaluated by industry, as reflected in the Code revisions described above.

Experience requirements for a variety of professions are developed on the basis of historical practice and subject matter expert judgment about how much time in practical field settings (e.g., internships, residencies, apprenticeships, etc.) is sufficient to develop skills at a level to work independently. While most of these practices and expert judgments are not documented in a form that allows reproduction of the reasoning underlying experience hours, long-standing codes and regulations reflect aggregate industry expert judgment. When new technology affords new training prospects—such as simulated UT exams or use of practice samples in lab exercises—it is important to recognize that operational experience provides exposure to a wide range of circumstances and develops adaptive skills that are only exercised in real-world settings. Over-reliance on simulations or lab exercises can lead to the development of “brittle” skills, which are less robust and flexible, i.e., do not easily translate to operations to address unexpected or emergent circumstances.

The relevant research to address the issue of time to develop knowledge and competency is found in the science of human learning and memory, and the associated application areas of training and expertise development. These research areas have been comprehensively reviewed in the attached appendices and provide the basis for the findings and recommendations described below.

2.0 Findings, Guidelines, and Recommendations

The overall findings of the literature review show that there are fundamental principles of learning and memory that can be applied to training and field experience. Factors that influence the rate of knowledge development and retention of material in learning research also affect learning in the applied settings of classroom and field experience. The integrative guidelines and recommendations provided below address (1) considerations that are general to all learning situations, (2) specific NDE circumstances, and (3) issues that concern regulatory oversight of training and experience requirements.

2.1 General Principles Applicable to All Learning and Training Situations

Research on human learning and memory, described in Appendix A, and training and instructional psychology research, described in Appendix B, has documented principles that apply across all learning circumstances. The conditions vary from the relatively artificial tasks conducted in laboratory studies, in which subjects learn various types of lists, to more realistic settings such as medical training and simulations of aircraft operations, in which students learn material that is integrated and applied. The principles and how they work are described below, along with examples and guidelines for general application.

1. **Total Time Principle**: As more time is spent learning, the chances of mastering a defined body of material increases, and a greater amount of material can be studied. With increased levels of study (more time, advanced course material), retention (memory) will be better. Time spent learning interacts with other factors, including spacing of practice, level of active learning, and knowledge testing.

   Guideline: Provide sufficient time for mastery of the specific body of material presented to students, within the overall structure of the instructional setting. An example would be providing learning time sufficient to demonstrate mastery of a calibration procedure.

   Research Technical Basis: Appendix A, Section A.5.1

2. **Forgetting Principle**: Forgetting occurs rapidly in the time period immediately following study, and levels off at 30%–50% of original learning. Repeated study at spaced intervals can reduce this effect, and enhances overall level of retention. Without continued practice, skills can decay within months following cessation of learning.

   Guideline: To establish higher levels of retention, material should be periodically re-studied at spaced intervals following mastery. Annual practice on lab samples is an example of reducing the effects of forgetting.

   Research Technical Basis: Appendix A, Sections A.5.1 and A.5.7; Appendix B, Section B.4

3. **Multiple Memory Modes Principle**: Everyday functioning involves multiple modes of memory: short-term (working), declarative memory for factual information, episodic memory for events, procedural memory for sequences, spatial memory for navigation, and perceptual memory for pattern recognition.

   Guideline: A variety of diverse situations should be arranged to selectively present material and engage the different forms of memory, based on the specific performance requirements of knowledge application. For NDE, the most appropriate example is field experience—here the trainee engages all types of memory.

Research Technical Basis: Appendix A, Section A.5.2

4. **Perceptual Learning Principle**: Experience can change how people perceive visual patterns, increasing the ability to extract information from the environment.

   Guideline: Provide practice viewing and classifying various types of patterns that may be seen in an operational environment. Laboratory practice with multiple specimen types engages perceptual learning.

   Research Technical Basis: Appendix A, Section A.5.2.2.1

5. **Incidental Learning Principle**: Learning can occur in the absence of overt behavioral change (e.g., test performance) as a result of exposure to material that is a secondary consequence of some other activity.

   Guideline: Provide opportunities for exposure and exploration of learning material that is not specifically devoted to test performance. Such activities might include evaluating plant operating experience (OE), case studies, and observing others perform procedures.

   Research Technical Basis: Appendix A, Section A.5.3

6. **Overlearning Principle**: Studying material beyond the point of mastery in close succession to the original study periods enhances test performance only if the test is relatively close in time to the learning period. Tests conducted at later points in time show diminished performance.

   Guideline: Continued practice after achieving mastery of material is only useful if the practice is spaced out in time (see Distributed Practice Principle, below). Learning material should thus be presented and tested to show mastery, but then practiced at longer intervals. Concentrated study prior to qualification testing involves overlearning, whereas annual practice does not.

   Research Technical Basis: Appendix A, Section A.5.4

7. **Distributed Practice Principle**: Learning time that is distributed across intervals rather than massed together yields better retention. Longer intervals between practice sessions are appropriate for longer retention intervals, such as several months between practice sessions for testing at one year after original learning. Varying study by using different versions of material or alternating types of problems can improve performance beyond simple repetition of the same material.

   Guideline: Develop instructional materials so that information that is important to retain is presented multiple times, with spacing of several days or weeks between repetitions. The intervals used will depend on the time available for training. Annual practice and use of site-specific samples are examples of distributed practice.

   Research Technical Basis: Appendix A, Sections A.5.4.2–A.5.4.5; Appendix B, Section B.3.1.1

8. **Feedback Principle**: Feedback that informs the learner as to the correct answer on a test, presented after the test is completed, enhances retention and can be used to correct errors. Feedback can also be used to help learners understand their process and to teach self-monitoring.

   Guideline: Provide feedback about correct answers on tests, with rationale (if feasible), soon after the test is graded. If appropriate, provide process and self-monitoring feedback. Comments and guidance by a supervisor are a form of feedback.

   Research Technical Basis: Appendix A, Section A.5.6; Appendix B, Section B.3.3

9. **Knowledge Testing Principle**: Testing can enhance retention, independent of further study.

   Guideline: Frequent testing during the course of learning, either through quizzes, classroom response systems, or self-testing, should be provided and occur without the opportunity for review. Material tested without re-study is remembered better than with additional study time. Recalling procedural steps from memory, such as calibration, is a form of knowledge testing that reinforces memory.

   Research Technical Basis: Appendix A, Section A.5.5; Appendix B, Section B.3.1.2

10. **Active Learning Principle**: Learning material that is processed more "deeply," i.e., with elaboration of content through meaningful engagement—as distinct from passive activity such as verbatim note-taking—will be retained better.

    Guideline: Use learning exercises that involve students developing explanations for concepts and other content, developing concept maps, and working with peers in discussion to explain material to each other. Field experience and laboratory practice both involve active learning.

    Research Technical Basis: Appendix B, Section B.3.4

11. **Multi-media Principle**: Learning is enhanced by using multiple sensory modes of presentation (visual diagrams, text, auditory) that can facilitate attention to relevant material, organization, and integration of the content with related material.

    Guideline: Integrate visual diagrams, text, and verbal explanations of learning material that lends itself to multiple modes of expression. Animations that show the relationship between weld flaws and waveform displays on a UT instrument incorporate multi-media.

    Research Technical Basis: Appendix B, Section B.3.6

12. **Transfer of Training Principle**: Practice on parts of tasks that are similar or identical to elements of the task being learned results in positive transfer of learning; i.e., the new task is performed better than if practice on related tasks had not occurred.

    Guideline: Use part or whole task simulation to provide practice on exercises that are similar to the objective of training or application (e.g., flying, examination of plant components). Combine simulation with operational tasks based on evidence from substitution studies and current regulation for aviation, which suggests a range of 10%–50% substitution of simulator time for operational time. Laboratory training with practice samples involves transfer of training.

    Research Technical Basis: Appendix A, Section A.5.8; Appendix B, Section B.3.7

13. **On-the-Job Training (OJT) Principle**: OJT is a systematic process based on analysis of the task to be performed, development of instructional situations from the operational environment, employment of instructional techniques such as demonstration and observation as warranted, and evaluation of performance based on increasing competence in operational tasks.

    Guideline: When training is to be conducted on the job, develop specific criteria for what is to be observed and then demonstrated by the trainee. Full demonstrations of tasks, as well as parts of tasks by instructors, should form the basis for student demonstration of proficiency.

    Research Technical Basis: Appendix B, Section B.3.2

14. **Proficiency and Expert Performance Principle**: Proficiency and eventual expertise develops over a long period of time, requiring deliberate practice of component tasks at increasing levels of difficulty to challenge the current skill level.

    Guideline: Establish entry-level proficiency in task-relevant environments, ensuring sufficient time to achieve and demonstrate competence. To increase skill levels to expert, provide increasingly challenging assignments and multiple opportunities to perform with assessment and feedback, over an extended multiple-year period.

    Research Technical Basis: Appendix B, Section B.3.8

15. **Self-assessment Principle**: Self-assessments of knowledge are inaccurate and tend to reinforce ineffective study and practice strategies such as re-reading.

    Guideline: Practice and refresher training should be based on objective assessments of knowledge and/or amount of time that has passed since specific knowledge or techniques have been applied. Annual practice should be based on time since certain procedures have been performed and anticipated application in the near future.

    Research Technical Basis: Appendix A, Section A.5.9

2.2 Specific Recommendations for NDE Training and Field Experience

Changes made in the 2011 Edition and Addenda of the ASME BPV Code, Section XI, Appendix VII for qualification as a Level II examiner reduced the experience hours required from 800 hours to 400 hours, with a minimum of 80 hours of field experience and 320 hours of laboratory practice. The U.S. Nuclear Regulatory Commission (NRC) has prohibited the use of the reduced experience hours in the 2011 Code, as described in the final rule (August 2017) (10 CFR 50 2018) because the effect of reducing operational experience is unknown. The following material addresses the implications of several of the general principles of learning discussed above and makes specific recommendations that pertain to experience qualifications for Level II examiners.

Several factors contribute to the motivation for the requested Code changes:

- Anticipated reduction in size of available NDE/UT workforce
- Lack of opportunities to detect and size flaws in operational plants because of their low base rate of occurrence
- Fewer opportunities for training during in-service inspections because of shorter outage periods and efforts to reduce occupational radiation exposure levels

These factors have led to a situation where trainees for Level II qualification gain some experience time on tasks such as making flow-accelerated corrosion (FAC) thickness measurements, but less experience performing other elements of a UT weld examination. Additionally, to increase the likelihood of successful PDI test performance, candidates are sometimes sent to EPRI to spend a week examining practice specimens prior to testing. Neither approach is optimal; making FAC measurements is a different task than UT weld examination, and concentrated practice before a test has been shown to be an ineffective strategy for long-term knowledge maintenance.

The following recommendations are meant to provide guidance to trainers and subject matter experts (SMEs) responsible for development of Code. As such, they reflect proposals already made—such as lab practice on flawed specimens, as well as issues that are more difficult to quantify but are of general concern—such as the intangible aspects of experiencing and observing conditions in operating nuclear power plants. The recommendations are:

- **Recommendation 1**: Increase opportunities to practice detection and sizing of flaws by use of laboratory practice exercises that simulate the types of examinations that would be done in the field. Provide sufficient time to ensure mastery of the relevant techniques.

  Applicable Learning Principles: Total Time, Transfer of Training, Perceptual Learning

  Discussion: This recommendation is consistent with proposed modifications to Code that would entail increased hours of lab practice on UT techniques and components likely to be encountered in field conditions. The amount of time provided for these exercises should be based on objective criteria for demonstrating mastery, such as proficiency ratings by supervisors, performance scores in practice tests, and PDI test pass rates. This recommendation does not imply that lab practice should *substitute* for field experience, but instead be used as an *augmentation*. Additionally, the objective definition of mastery can be somewhat circular, i.e., passing a practice test or PDI test that is similar to lab exercises does not necessarily ensure effective performance in the field. Thus, any implementation of a lab practice regimen should be carefully monitored with multiple sources of data—Level III ratings of actual field performance, pass rates on PDI testing, and (potentially) “check examiner” oversight of newly credentialed examiners by more experienced personnel. This latter concept would be similar to use of a “check pilot” in aviation who accompanies trainees during flight testing to ensure correct performance and adherence to procedures. Potential “provisional” Level II certifications may be warranted until sufficient data is gathered to address criteria for skill mastery and effective performance in operational settings.

- **Recommendation 2**: Define the relationship between hours of experience and frequency of UT task performance. Current proposals for lab practice and field experience hours use a table of frequency X hours per task to yield total experience time. Establishing a more detailed basis for what tasks and subtasks contribute to the time estimates will provide employers and regulators with a better understanding of how to implement and evaluate Code modification proposals.

  Applicable Learning Principles: Total Time, OJT, Distributed Practice

  Discussion: A problem with basing qualifications on experience time is the specific definition of “time.” While time and frequency of task performance are related, it is unclear whether the number of times the task is performed, or the duration of the task performance, is the key variable in developing proficiency. In this case, the principle of distributed practice is likely to be more applicable, in that the frequency and variability of the circumstances under which examinations are performed will contribute to proficiency development more so than simple total time. For example, performing three 20-minute examinations in different plants is a better training experience than one 60-minute examination in a single plant.

  The ASME NDE (ANDE) working group initiative attempted to develop a task proficiency-based approach to training through the development of “job cards,” which could be used to track experience and examination procedures performed. However, ANDE does not have an objective basis for determining how many such examinations need to be performed to achieve Level II proficiency. This basis will require systematic estimates by SMEs, as well as review of pass-rate data for PDI.

- **Recommendation 3**: Provide sufficient field experience in nuclear plant in-service inspections to develop proficiency of the knowledge and skills necessary to conduct examinations in realistic operational settings. The amount of time in field experience should complement laboratory practice such that total experience time is equivalent with the Code currently accepted by the U.S. Nuclear Regulatory Commission (NRC) (800 hours). This would thus entail more hours of operational field experience than is currently proposed in Code modifications, but less than currently required by recognizing the value of lab practice. Field experience should be based on a structured approach to OJT to ensure that all techniques addressed in laboratory practice are observed and demonstrated in the field under supervision, and feedback provided to the trainee concerning their performance.

  Applicable Learning Principles: Total Time, Incidental Learning, Distributed Practice, OJT, Proficiency and Expertise, Active Learning.

  Discussion: This recommendation could reduce the field experience hours substantially, recognizing the potential contribution of laboratory practice. Although the skill of detection and sizing of flaws is usually not exercised during in-service inspections because of the low rate of flaw occurrence, there are many other aspects of developing proficiency for skill application in operational environments. These aspects include learning how to navigate, executing procedures in complex circumstances, observing multiple geometries, addressing issues of coverage limitations, and generally learning flexibility to adapt the examination process while maintaining compliance with the qualified procedure, as necessary. It also provides the trainee with experience in diverse operational cultures and expectations.

  The Code modification of 80 hours of field experience appears to be minimal—that amount of time could be obtained during a single, extended refueling outage period, and would not yield the diversity of settings and procedure application that field experience is designed to impart. Because proficiency and eventual expertise has been shown to require as much as 10,000 hours to develop, reducing the field experience hours to the amount proposed would likely result in trainees being good at detecting flaws in practice samples, but quite naive with respect to operations, and thus not prepared to operate on their own—which is one of the main capabilities expected of Level II examiners. As with the laboratory practice recommendation described above, this recommendation for field experience may be implemented on a "provisional" basis, with the option for fully implementing it should data suggest its effectiveness.

- **Recommendation 4**: Develop a schedule for laboratory practice that distributes the practice sessions for various techniques over intervals separated in time, e.g., a week or a month, based on what is feasible for trainees, oversight supervisors, and lab facilities. Incorporate testing of previously practiced material.

  Applicable Learning Principles: Distributed Practice, Knowledge Testing

  Discussion: This recommendation is intended to enhance the effectiveness of laboratory practice by incorporating distributed repetition, i.e., spacing of lab exercises over time so that more durable learning takes place. The current approaches to preparation for PDI involve massed practice on samples in the time period immediately preceding the test. This approach, which is similar to "cramming," involves repeated scanning using techniques and samples similar to what will occur on the qualification test. This can lead to overlearning, i.e., practice beyond the point of mastery—and may enhance test performance, but will not yield learning that is as durable as if practice were spaced over intervals of time. For example, rather than doing all of the practice in a single week prior to qualification testing, preparation 6 weeks in advance, with several hours of testing on alternate days, would incorporate the benefits of distributed practice. The use of testing of previously practiced samples is meant to reinforce what was learned in prior practice.

The feasibility of distributed practice is an issue that needs to be considered, particularly if most practice opportunities involve travel to a location where samples are available for practice. Development of simulated samples with embedded flaws based on the samples available at EPRI would allow practice to be carried out at the trainee's home location and scheduled in such a way as to distribute the exercises over time.

- **Recommendation 5**: Develop a protocol for laboratory practice that incorporates feedback to the trainee concerning their performance, addressing (1) performance accuracy, (2) process of examination, (3) how the trainee can monitor their own performance, and (4) increasingly difficult problems to challenge the current trainee skill level.

  Applicable Learning Principles: Feedback, Proficiency and Expertise

  Discussion: Implementation of greater amounts of laboratory practice for UT procedures permits the opportunity to provide more feedback concerning performance. Lack of performance feedback—other than "pass/no pass"—is a common concern raised by examiners. More extensive lab practice can be used as a basis to provide feedback not only about performance accuracy—whether the flaw was correctly detected and sized—but also about how the trainee performed the task, and how they can watch their own performance to ensure proper task completion. This level of feedback is likely to be more intensive for early trainees who will require considerable oversight, while in the latter periods of lab practice, less scrutiny would be necessary. Additionally, the feedback provided by Level III supervisors might address how such examination procedures would be performed in the field and practical considerations that are typically encountered in operational settings. Incorporation of graduated levels of difficulty for the practice specimens will permit the trainer to challenge the current skill level of the trainee, so that current knowledge is pushed and extended.

2.3 Summary

The NRC is concerned with ensuring that NDE/UT examiners attain proficiency in their job skills to perform effectively on the job. The issue of experience hours is not unique to the nuclear industry; supervised time requirements exist for many skill-based jobs, including aircraft piloting, nursing, medical practice, and numerous other occupations. The time requirements exist in order to ensure that trainees have sufficient opportunity to experience and perform their job tasks at an acceptable level of proficiency so that they can work unsupervised—the "journeyman" level of performance. The difficulty with NDE Level II experience is defining exactly what "proficiency" means; currently, passing the PDI qualification test after the required amount of field experience is often the accepted definition.

Proposals to change one of the core aspects of proficiency development—e.g., field experience hours—should be carefully evaluated and addressed conservatively. The scientific literature on human learning suggests that more time on job-relevant training is better. Thus, the proposals to increase laboratory practice are commensurate with guidance from the literature. Reducing field experience, however, is a more problematic issue. The current proposal to reduce field experience to 80 hours would seem to provide only a token amount of time. More appropriate would be to ensure that field experience actually does provide experience in observing and executing UT examinations under supervision, and that this is done frequently enough to develop operational proficiency. There is latitude to reduce the overall number of experience hours with a corresponding increase in lab practice. However, the proposals should be evidence-based with a rationale for accumulating adequate experience to develop proficient examiners, rather than simply being able to pass a single demonstration test.

10 CFR 50. 2018. "Incorporation by Reference of American Society of Mechanical Engineers Codes and Code Cases." *Code of Federal Regulations*, U.S. Nuclear Regulatory Commission, Washington, D.C. Available at https://www.nrc.gov/docs/ML1613/ML16130A530.pdf.

Appendix A – Literature Review of Human Learning and Memory

Summary

This appendix reviews research from the science of human learning and memory as it pertains to the core issues in NDE training—amount of time spent in the field to develop skill, the progressive loss of skill over periods of disuse, and the importance of specific learning factors such as practice, testing, and feedback. Additionally, issues related to very long-term retention, transfer of training, and metacognition provide a basis for interpreting current practices in NDE training. The research reviewed yielded a number of general principles that are applicable to NDE training, which are described below.

The earliest studies of human learning and memory and research up to the present day confirm that learning takes time, that forgetting is a simultaneous process, and that balancing these two processes can help to enhance retention of learned material. Thus, the total time hypothesis of learning is still the best principle to use when considering the structure of training curricula, i.e., **more time spent learning will result in better and longer retention of information**. This principle applies in very short retention intervals as well as over multiple decades. Numerous intervening factors can facilitate how learning time is spent, which can also be incorporated in training implementation.

Multiple types of memory are used by NDE examiners, including short-term working memory, long-term declarative memory for factual information, long-term episodic memory for events, procedural memory for sequences, spatial memory for plant layouts, and perceptual-representational memory for specific patterns observed in the past. With the development of expertise through experience, working memory can grow in capacity so that more senior personnel can hold more relevant information about an inspection in their immediate awareness. The inter-relationships between these types of memory and training for NDE in operations is complex; the key principle resulting from this research is that **multiple modes of learning and memory contribute to acquiring, maintaining and applying NDE skill, and that operations experience and classroom laboratory learning are complementary rather than interchangeable.** Thus, understanding the multiple types of memory and their high-level functions in the context of NDE examination and training can be useful in considering such factors as how to organize examination tasks, the design of procedures, the sequence and hierarchy of teaching physics and acoustics, and the use of personal experiences to illustrate specific examination findings.

Measurement of knowledge and skill by performance tests does not completely represent what is known. Simple exposure and exploration within an environment (latent learning) or manipulating information for some other purpose (incidental learning) can result in task-relevant knowledge that can be applied when conditions warrant. Specific relevance of this general finding concerns how experience hours are gained by NDE trainees—learning may be taking place simply by helping a more experienced examiner. The general principle derived from this research is that **learning occurs in the absence of specific performance demonstrations or requirements, and that performance testing gives an incomplete portrait of the extent of knowledge.** These findings suggest that in training NDE examiners, there is an impact of the experience obtained by junior personnel as they acquire experience hours. This can include developing a “cognitive map” of complex work environments, heuristics for how to scan various structures, and mental models of procedure execution. Specific knowledge and performance

tests for this type of latent and incidental learning do not exist for NDE field experience, but interviews with examiners suggest its importance ("the first couple of years were a blur… it takes time to know where you are going") (Sanquist et al. 2018).

Overlearning—study of material beyond the point of mastery in close succession to the original study period—is a common characteristic of practical education and training. Research has shown that when the test period is relatively close in time to the learning events, performance is enhanced in contrast to learning to simple mastery, i.e., one perfect test. However, if testing is conducted at later points in time, the effects of overlearning are much diminished. There is some evidence that motor skills, such as equipment assembly/disassembly, show longer-term positive effects from overlearning. Conceptual and verbal material, however, tend to degrade over time despite overlearning. Research indicates that overlearning across separate sessions leads to enhanced performance; by spreading out the study time that leads to mastery, retention is enhanced in later performance tests. A more analytic view of overlearning suggests that the underlying principle is that **short-term overlearning results in better immediate test performance but is not durable. Overlearning associated with repeated exposure to foundational material to learn more complex concepts and operations entails spacing of repetition and leads to more durable learning for the long term.**

Studies of repeated learning and practice effects in humans have focused on the distribution and variability of the practice sessions. Many years of parametric study of the spacing effect in human learning suggests that there is an interaction between spacing intervals and retention intervals; the beneficial impact of spacing depends on the length of time between learning and test. Further, varying how study is performed by using differing versions of the material or types of problems—known as *interleaving*—can improve performance beyond simply repeating the same material that has been presented before. The general principle resulting from this work is that **practice distribution and variability leads to more durable learning than concentrated practice on identical material; longer intervals between practice sessions are appropriate for longer retention intervals, such as several months between practice sessions for testing at a one-year retention interval.** For NDE training, if the eight required hours of annually required practice are done all in one sitting, e.g., with practice samples, there will be much less retention of knowledge than if practice were performed in two-hour intervals four times throughout the year. Similarly, interleaving different types of samples and welds would capitalize on the beneficial effects of practice variation.

The "testing effect" has been known for many years in educational settings, and laboratory studies have revealed that while repeated study is more beneficial for immediate test performance, testing—with or without feedback—produces better retention over longer periods. Related findings indicate that when learners engage in self-explanation of relationships in studied material or questions are asked regarding relationships, recall is better. In general, **testing of previously learned material is preferred over simple re-study, and testing that involves elaborating semantic linkages within the material can potentiate retention.**

The concept of feedback in learning can be traced to reinforcement, a consequence of behavior that increases its likelihood. In cognitive tasks, feedback is generally extrinsic and can be delivered either immediately or at some delay. Further, feedback can provide varying levels of information, such as correct-incorrect, hints regarding the correct answer (known as scaffolded feedback), and correct answer. Voluminous studies in a wide range of tasks show gradations in performance depending on how and when feedback is provided. It is generally the case in both laboratory and traditional classroom settings that **retention of information is best when delayed correct answer feedback is provided.** The delay in laboratory setting is typically provided by structuring the learning and test sessions so that feedback is provided after all items have been tested, while delay of a day or more shows beneficial effects in classroom settings. Studies of error generation during learning suggest that the general philosophy that errors should be minimized during learning is both unrealistic and incorrect—instead, **errors represent learning opportunities when paired with appropriate feedback and can specifically enhance learning.**

A consistent issue in our field interviews with NDE examiners is their concern with the lack of feedback during performance testing for various procedure qualifications. They are simply told whether or not they passed. Senior NDE examiners have failed these examinations, as well as more entry-level personnel, and they have little information available to them to understand what they did wrong and to correct their errors. The testing organization is concerned that feedback will compromise the test items, which are limited in number and expensive to produce. This understandable concern might be mitigated somewhat with a scaffolded feedback approach employed during the test examinations, e.g., providing cues without giving away the answer. Alternatively, post-test debriefing using recordings from the examination instrument might represent a method of providing feedback beyond pass/no-pass, while maintaining integrity of specific test samples. Similarly, use of a low- or medium-fidelity simulator could allow generation of an unlimited number of weld flaw variations, with concomitant performance feedback.

The few studies of very long-term retention, i.e., multiple decades, have used cross-sectional research methods to assess knowledge maintenance over long periods of time. The general finding from these studies is that **knowledge retention declines over the first five years after learning, then stabilizes at levels around 40% for as long as 35 years**. Correlation of very long-term retention with original learning factors such as grades received and level of training suggest that **better performance in original learning and greater levels of advanced training lead to better long-term knowledge maintenance.** Further laboratory studies using shorter retention intervals suggest that periodic relearning for “marginal knowledge” serves a preventive maintenance function and increases likelihood of retention for longer periods. Specific **studies of optimizing relearning schedules suggest that providing additional exposure to material at intervals where knowledge has decayed somewhat, but prior to final testing, is beneficial for longer-term performance.** This general pattern of results has implications for continuing training in NDE. First, the studies suggest that longer periods of training, i.e., continuing to advanced levels, lead to higher overall levels of retention. Thus, a general training strategy for continuing education should emphasize progression from basic to more advanced topics and/or procedure complexity. In practice, this means that individuals attaining Level II certification should not only maintain that certification but should study toward Level III—this will help to consolidate their more basic knowledge.

Transfer of training occurs when what is originally learned is applied in the performance of a different task. This is essentially a generalization of prior learning to a situation with some degree of similarity to the original. Transfer can be positive, i.e., the new task is performed better than if no original learning had taken place, or negative, i.e., the new task is performed more poorly. Substantial research has been conducted to determine the relationship between original learning and subsequent application. The general finding is that **to the extent that there is similarity between the new task and original learning, positive transfer occurs. As the degree of similarity differs, or if the new task involves opposite response patterns from what was learned, negative transfer occurs.** Studies of training transfer are particularly relevant to issues of appropriate experience for professional qualification and the use of simulators in developing skills. Knowledge transfer is an issue that was reflected in the task analysis study of manual ultrasonic inspection. The concern was expressed that entry-level personnel were not provided with adequate field experience opportunities—specifically, that performing flow accelerated corrosion inspections was not appropriate experience for understanding and applying manual UT for weld inspections. More generally, the progression from basic principles in classroom training to use of that knowledge in evaluating potential flaws is a classic problem in knowledge transfer.

A key set of factors that influence learning and retention are individual characteristics, such as self-assessments of what you know and how well you know it. The general findings from studies in this area are that **self-assessments of knowledge are inaccurate and tend to reinforce ineffective learning strategies such as re-reading.** People will tend to avoid using learning strategies such as spaced practice and self-testing in favor of intuitively appealing and relatively easier strategies such as review and re-reading. Research indicates that **"desirable difficulties" in learning that may entail more mistakes in the short run will pay dividends of better retention in the long run.**

With respect to NDE training, the principles summarized above suggest several approaches, albeit somewhat idealized, to implementation. These would include:

- A programmed learning approach to classroom/text-based material that uses appropriately spaced and elaborated repetitions, as well as testing with feedback provided immediately following test completion. Standardization of this material is recommended to reduce variation across training providers.
- Alternating classroom/text-based material so that subject-matter variety is provided.
- A period of field experience commensurate with observing and assisting with the types of exams to be performed upon certification.
- Practice in summarizing findings from field experience assistance to develop expert working memory capability.
- Performance demonstration tests that focus on more than a single procedure or weld type to promote variability in practice and retrieval-based testing.
- Development of a means to provide some type of feedback for performance testing that is more informative than pass/no-pass.
- Development of a catalogue of common errors and feedback that is focused on correcting inspection process elements that engender those errors.
- Requirement for refresher training that is based on the expected intervals between knowledge use and expected decay.
- Requirement for testing refresher training, possibly developed in a manner so that periodic testing/training can be performed concurrently.

A.1 Introduction and Background

Certification as an nondestructive examination (NDE) examiner is a process based upon training concerning materials, physics, acoustics and other relevant disciplines, as well as obtaining a prescribed number of hours of experience in industrial examination settings. Certification is a combined process of (1) obtaining the necessary experience and classroom training, and (2) demonstrating proficiency in a practical demonstration of performance in the Performance Demonstration Initiative (PDI) setting. In the US, these requirements are established by American Society of Mechanical Engineers (ASME) Code Section XI, Appendices VII (training and experience hours) and VIII (performance demonstration).

The certification body for training and experience hours is the employer or training institution, and the Electric Power Research Institute (EPRI) provides performance demonstration proficiency testing (NRC 2009). The various requirements associated with these processes have been the subject of proposed changes, such as a reduction in the number of experience hours, substitution of laboratory for training in lieu of field experience, and newly-developed certification processes based on job cards detailing specific examination experience (Cumblidge 2018).

Recent editions of the ASME Code have reduced the practical job experience hours for personnel certification. The 2011 revision to the ASME Appendix VII Code, for example, reduced the required experience hours for Level II certification from 800 to 400, with 320 hours comprising laboratory training (Cumblidge 2018). This is essentially a 90% reduction in the number of field experience hours from earlier code requirements. The U.S. Nuclear Regulatory Commission (NRC) has prohibited the use of the reduced experience hours in the 2011 Code, as described in the following paragraph from the final rule published in August 2017:

> In the 2011 Addenda and 2013 Edition, the ASME BPV Code added an accelerated Appendix VII training process for certification of ultrasonic examination personnel based on training and prior experience, and separated the Appendix VII training requirements from the Appendix VIII qualification requirements. These new ASME BPV Code provisions will provide personnel in training with less experience and exposure to representative flaws in representative materials and configurations common to operating nuclear power plants, and they would permit personnel with prior non-nuclear ultrasonic examination experience to qualify for examinations in nuclear power plants without exposure to the variety of defects, examination conditions, components, and regulations common to operating nuclear power plants.
>
> The impact of reduced training and nuclear power plant familiarization is unknown. The ASME BPV Code supplants training hours and field experience without a technical basis, minimum defined training criteria, process details, or standardization. (10 CFR 50 2018)

The revised Code has been prohibited for use by the NRC in the Code of Federal Regulations (10CFR50.55a(b)(2)(xviii)(D).

Annual practice is specified by ASME Code Appendix VII, Section 4240, such that “Personnel shall practice UT techniques by examining or by analyzing prerecorded data from material or welds containing flaws similar to those that may be encountered during inservice examinations.

This practice shall be at least 8 hr per year and shall be administered by an NDE Instructor or Level III. No examination is required."

PNNL has conducted human factors research to address the technical basis for more rigorously developing evaluation training and certification requirements for NDE examiners.

Development of a sufficient technical and procedural knowledge base to qualify for admittance to various occupations is a broad issue faced by many professions. Practice domains as diverse as surgery, radiographic image interpretation, aircraft and maritime piloting, as well as NDE entail both practical experience and technical knowledge requirements. For any such job, there are required criteria for entry and maintenance of certification. The original establishment of these criteria is often based on a combination of historical practice, expert practitioner judgment, and practical considerations of the time and cost to develop and maintain expertise. Specific mastery of domain material is also required and generally assessed by the conventional approach of knowledge tests and, in some cases, practical demonstrations.

To provide a more empirically-based foundation for establishing and maintaining job-specific competencies, it is necessary to draw from the literature concerning human learning, memory and training (the applied focus of learning). This is an extremely voluminous research domain that has been the subject of study for over 140 years (Mayer 2012). Considerable progress has been made during that time in elucidating some of the basic mechanisms of how people learn, retain and apply information. There has been less progress in the specific application of this knowledge to the practice of education and training, but core principles do exist to guide specific program development and student evaluation. This appendix selectively reviews the research literature pertinent to the following general questions:

- How much time does it take for a person to learn to a criterion of mastery?
- What are the factors that determine retention of information (e.g., time, repeated use, practice, etc.)?

These basic research questions pertain to the more applied focus of concern for developing a technical basis related to NDE training, which entails the following questions:

- How many hours are required to become proficient?
- What kinds of hours are necessary to become proficient (field, classroom, lab)?
- What kinds of retraining or practice are necessary, and over what period?
- How much retraining or practice is necessary to maintain and/or enhance competency?

Our goal in this appendix is to review the general principles of human learning that have been identified in basic research studies, practical applications of those principles in training and instructional settings, and the implications for current and future approaches to NDE training. To provide boundaries for research literature with such a long history and multiple perspectives, we will use the lens of NDE training to focus on research that is specifically relevant. Characteristics of the NDE job and our understanding of current NDE training will be used to select, interpret and apply relevant information from the research literature. NDE job characteristics and training include:

- Factual knowledge requirements of physics and other engineering sciences
- Demonstration of knowledge by traditional question-answer tests

- Demonstration of procedural skill on samples
- Classroom or other study-based instruction on relevant physical science
- Field-experience-based learning in appropriate job settings.

Our goal is to provide a representation of the current theory and psychological mechanisms applicable to NDE training and assessment and to translate this material into specific guidance, as appropriate. In this sense, the findings presented in this paper should be considered an “interpretive review” of contemporary and historical research that is specifically germane to the technical basis questions listed above.

A.2 The Context: Technical Skills Training in the US

This section presents an overview of a recent study by the National Research Council (2017), which identified the complexities of skilled workforce development in the US. As shown in Figure A.1, during secondary education students may pursue either an academic or career and technical education (CTE) track. Subsequently, skills can be developed in a variety of postsecondary settings.

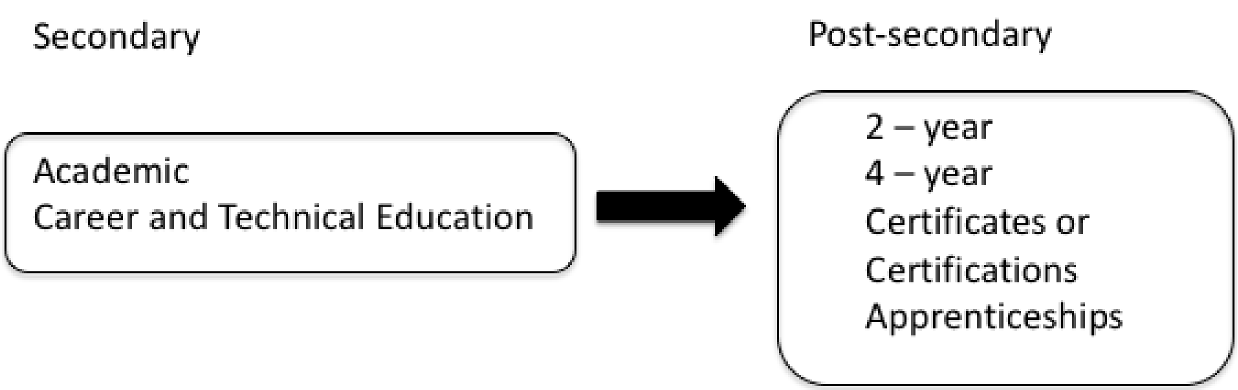

Figure A.1. Technical education pathways in the US. CTE—Career and Technical Education. (based on National Research Council 2017).

This model, while descriptive of available training pathways, is not a formal, nationwide system that proceeds in linear fashion. Instead, while these pathways are available to students, “there is no single, formal system for training workers, and Americans are responsible for finding their own path, based on their preferences, capacities and means” (National Research Council 2017, p. 63).

Technical training “tracks” begin in secondary education settings through course work in CTE subjects. Concentration in these types of courses has declined from 33 percent in 1982 to 19 percent in 2009 (National Research Council 2017). Further, the skill mix in CTE has shifted away from what might be considered “trades” such as manufacturing and engineering technologies, communications, health care, culinary and consumer services, which seems to reflect the overall decline in manufacturing in the US and the corresponding rise of the service economy. Entering NDE as a result of secondary CTE is an unlikely pathway, given these findings; further, the results of the NRC Manual Ultrasonic Testing (UT) Task Analysis study (Sanquist et al. 2018) suggest that many examiners find their way into the profession through family connections and vendor-based training.

Postsecondary education provides several potential routes for technical skill training, including traditional 4-year colleges, 2-year colleges, certificate or certification programs, and apprenticeships. The American Society of Nondestructive Testing (ASNT) maintains a list of educational organizations providing various levels of nondestructive testing (NDT) training, varying from full BS and MS degree programs in the relevant engineering sciences to corporate-based training. The wide extent of offerings in these organizations and the different constituencies they serve make any sort of classification or evaluation difficult.

Apprenticeships are “formally structured work training programs in which workers are employed in a job and earn wages while receiving on-the-job training and technical instruction” (National Research Council 2017, p. 73). Current data suggest that less than 2 percent of US workers participate in such programs. Internet searches for “NDT apprenticeship” identified two formal NDT apprenticeship programs based in the United Kingdom and none in the US. However, NDT vendors do hire “Level I” examiners who are essentially trainees, and in-house or external technical instruction may provide a similar function to apprenticeship.

Certification falls into the category of “alternative credentialing.” Certification is essentially a document, awarded by a certification body, attesting that the holder has demonstrated knowledge and skill through an examination. In the case of NDT, the certification is focused on the content guidelines of the ASNT (and that training can be provided by the employer or one of the other postsecondary tracks described above). For NDT, the certification body is the employer or training provider.

Quality control across the various technical skill training tracks varies considerably. In the case of 2- and 4-year public and private non-profit educational institutions, a variety of federal, state and regional accreditation bodies ensures a degree of uniformity. For-profit educational institutions—which have doubled in number between 2000 and 2015—are not accredited in a similar way, and the quality varies considerably; it is noteworthy that the completion rate of students at for-profit schools is 27 percent, compared to 65 percent and 58 percent at private non-profit and public institutions, respectively. Apprenticeships in the US are not tied to a competency-based standard (in contrast with European programs). Certificate programs are also not linked to specific accountability standards, making evaluation of individual certificate holders difficult.

A.3 Learning Myths

Myths are generally defined as fictitious or unscientific accounts or beliefs. In the anthropology of folklore, myths are considered narratives that structure beliefs about how the world works and are shared among specific groups of people, such as students and educators (Diamond and Moezzi 2000). Myths often have both true and false elements. Thus, whether true or not, myths are a way of communicating long-standing, commonly-held beliefs. In the area of learning, there are a variety of myths that are often translated into specific instructional and study approaches (Dunlosky et al. 2013). These include:

- Test performance is equivalent to learning
- Concentrated and repeated study over a short period is the best way to learn
- Total time spent studying is the most important element in learning
- Studying in the same place consistently, i.e., keeping conditions constant, enhances memory

- Immediate feedback is important for learning
- Re-reading is better for retaining information than testing yourself
- Highlighting is an effective learning technique.

Much of the contemporary research in human learning and studies of training reviewed in this appendix suggests many of these beliefs are incorrect, or at least only applicable in very narrow circumstances. While the application of research findings is sometimes difficult in practical education settings, broad application of unsubstantiated beliefs about learning can be counterproductive.

A.4 Models of Learning and Human Information Processing

Learning is defined as a relatively permanent change in behavior resulting from experience, distinct from changes in motivation or maturation (Gagne 1965). The “experience” aspect of this definition may be a simple sensory stimulus such as a flash of light, or as complex as a multi-year academic and practical medical training curriculum. Cellular changes in the brain result from experience, and specific neural mechanisms underlie learning and memory (Gerbier and Toppino 2015).

In considering the voluminous research concerning human learning in various contexts, it is helpful to use models to provide focus and organization. For the situation we are considering here, i.e., training in a technical field that involves considerable science and engineering material as well as practical skill development, Figure A.2 captures the essential elements. There is an initial period of instruction, during which learning takes place. This is followed by assessment to determine the extent of the learning, i.e., is it sufficient to consider that the student has mastered the material? Following learning assessment, there is application of that knowledge in the job environment and usually an ongoing process of refresher training.

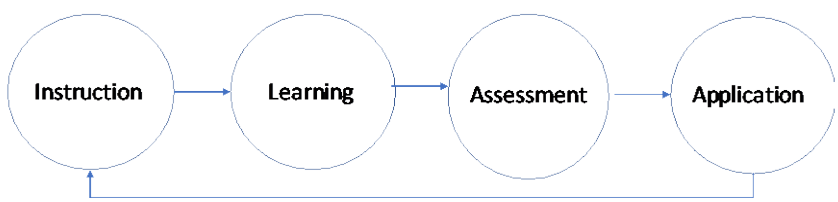

Figure A.2. Basic model of the technical training process (adapted from Mayer 2011).

The model shown in Figure A.2 is useful because it emphasizes the linkages between external processes, i.e., instruction and assessment, with internal processes, i.e., learning and application. Instruction and assessment are influenced primarily by the organizational structure in which they take place and to some extent by the individual instructor. Learning and application are processes that are dependent on the individual person; further, as will be discussed below, assessment—although meant to be a measure of learning—does not necessarily reflect the ability or likelihood of proper information application imparted during instruction. These distinctions are important because the present system of instruction and assessment for NDE examiners is fragmented with varying approaches and quality (Stephens 2014), and assessment via performance demonstration does not necessarily translate to skilled field application.

A more technical view of the individual learner is portrayed in the model shown in Figure A.3—one of the early “information processing” models of cognition.

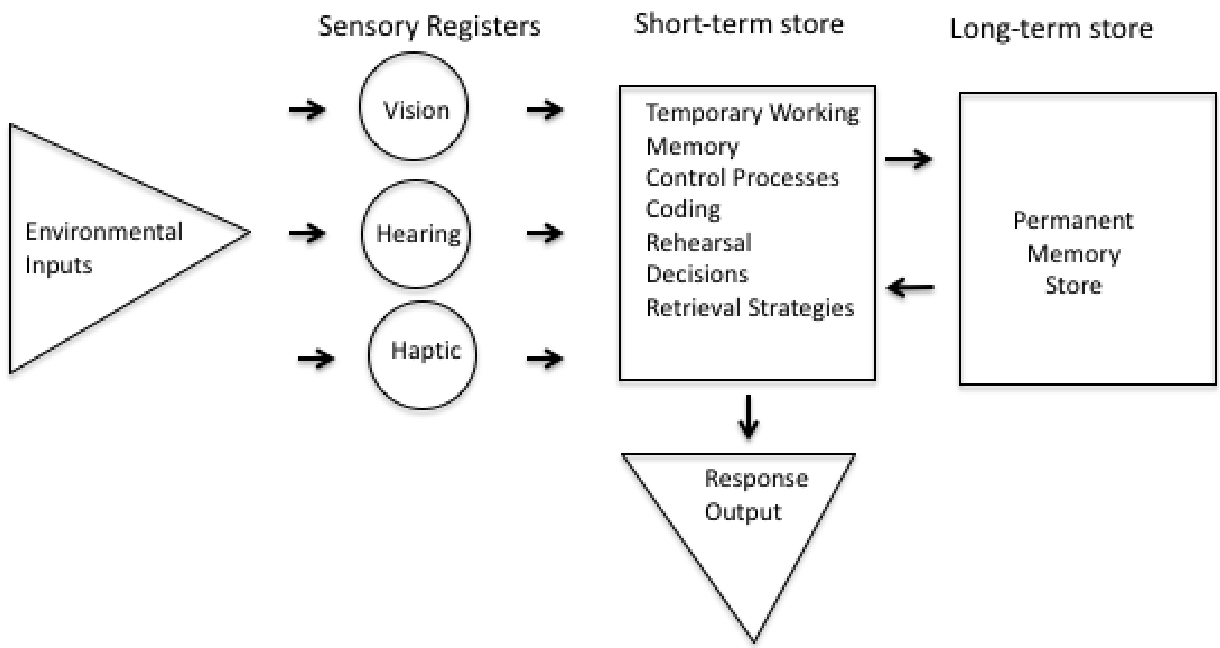

Figure A.3. Flow of information through memory system (after Atkinson and Shiffrin 1968).

In the information processing model, a learning situation would be based on environmental input presented to the individual by means of various sensory channels. This information is perceived and attended to via a working memory (basically, the current contents of conscious thought), and stored in long-term (permanent) memory. Various situational stimuli result in “retrieval strategies” that call information from the permanent memory store and produce a response that is appropriate to the circumstances. The basic learning model (Figure A.2) and the information processing model (Figure A.3) are useful for introducing the various perspectives from psychological research that need to be considered in a review of this type.

A.4.1 Measurement and Methods

The study of learning has used human and animal subjects in a variety of contexts, ranging from highly-controlled laboratory situations to naturalistic settings. The search for general laws or principles of learning has emphasized the use of structured situations for purposes of replicability and comparison. While learning, retention, and knowledge application are lifelong processes, researchers tend to study more tractable time spans. Longitudinal studies are relatively rare, although some researchers (e.g., Bahrick 2000) have made credible analyses of knowledge retention over decades.

Experimental laboratory studies of human learning and memory have focused upon verbal material or cognitive-perceptual-motor skills. The former type of study typically employs lists of words or other verbal material to examine the effects of various conditions of study on later ability to recall or recognize the material. Research on cognitive-perceptual-motor skills tends to focus on the acquisition of various levels of skill as measured by speed and accuracy of performance over time. Skill covers a broad range of tasks that focus on accurate output, such as telegraphy, tracking, and complex pattern reproduction, but in terms of learning, the processes are similar (Fitts 1964; Adams 1987).

Training, instructional and educational psychology studies are concerned with learning, retention and application of integrated material—either in relation to specific job/task performance (training), or as part of a curriculum of general education (instructional and educational psychology). The concerns of both areas are the same—how much is learned, over what period of time, and what is retained? Methods employed in this domain tend to involve standard tests of knowledge or task performance in which students are assessed following the training or instruction period.

Learning studies with non-human species are less relevant to the concerns of this review, but sometimes there are notable similarities between findings in animal and human studies that suggest general underlying mechanisms. Animal studies have used a variety of learning settings, including mazes, puzzle boxes, object and pattern discrimination, and physiological conditioning. The early studies by Thorndike of cats escaping from puzzle boxes were fundamental in showing that the effect of behavior was a key element in learning (later referred to as reinforcement or feedback). Thus, when appropriate, this review will illustrate general principles that apply across species.

Quantifying learning, retention and application of information (transfer of training) employs a number of standard measures, including:

- Time to criterion performance (e.g., mastery)
- Percent correct performance
- Percent information retained over time
- Percent errors
- Performance improvement over time

These classes of measure are manifested in various ways according to the specific focus of the study, e.g., laboratory studies of list learning, training of job/task performance, classroom learning, and animal studies. The common threads are time and accuracy.

A.4.2 Psychological Research Perspectives

For this review, we will address research literature from four areas that are related by their concerns with learning and skill development but distinguished by the underlying problems addressed, the theories, and the methods. These four domains are: (1) human learning and memory, (2) training and instructional psychology, (3) expert performance in specific domains, and (4) domain-specific studies in areas such as medicine, aviation and other job applications. This appendix discusses the basic science of human learning and memory; Appendix B discusses training, instructional psychology, and studies of expertise.

The fundamental science underlying training is that of human learning and memory. The study of human learning and memory has a 140-year history and has evolved through a variety of theoretical and empirical approaches. Mayer (2011) discusses the interactions between what he terms the Science of Learning (SOL)—basic research—and the Science of Instruction (SOI)—applied research and practice. Unfortunately, for most of the 20th century the SOL was essentially a one-way street in which research was conducted with people in highly-controlled but very contrived and unrealistic settings, such as memorizing lists of nonsense syllables. Simultaneously, the SOI created various instructional principles based more on philosophy than data, and there was little cross-talk between the disciplines.

This situation changed in the 1990s when various factors combined to foster more interaction. These factors include SOL reaching a theoretical and empirical “dead end” wherein further refinements of theory based on artificial situations yielded diminishing returns, a turn by cognitive researchers toward neuroscience work, increasing cross-talk between neuropsychologists and traditional experimentalists, and a challenge from educators to learning theorists to address learning in real-world tasks. Some of this latter impetus came from the armed services, which had a considerable short-fall of qualified recruits during the transition to an all-volunteer force.

Training can be considered the *applied* aspect of human learning and takes a variety of forms, including traditional formal discipline (study-test), situated approaches such as on-the-job and apprenticeship, and use of simulators. Training is generally considered more focused on specific technical subject matter for a job or task, as distinct from *education*, which implies more general knowledge associated with broad subject areas, such as history or physics (Healy et al. 2012). Considerable work has been carried out by the military, particularly since WWII (Gagne 1962). Related work has been conducted in the area of *instructional* or *educational* psychology, with a focus on more traditional classroom and subject-matter learning.

The study of expertise is a somewhat distinct sub-field in that it combines the concerns of human learning and training focused on domain-specific mastery and maintenance of skill through practice. Early studies of expertise in telegraphy (Bryan and Harter 1899) illustrated various phases of learning to mastery. In the 20th century, the studies of chess players by (de Groot 1946) eventually led to a focus on expert performance (Chase and Simon 1973), with subsequent studies across many domains (e.g., Ericsson 2009). The studies of expertise tend to be focused upon specific factors, such as long and deliberate practice, that lead to superior performance.

Finally, a variety of domain-oriented studies in areas such as surgical skill, aircraft piloting proficiency and other jobs have been conducted to assess the utility of various approaches to learning in specific fields. Examples of these include means by which elements of specific jobs or professions, such as surgical procedures, sonar operation, maritime navigation or aircraft piloting are trained and evaluated. This fourth area will be concerned with several jobs that have field experience requirements similar to that of NDE.

A.5 Human Learning and Memory

This volume of the appendix discusses research from the perspective of laboratory experimental studies. This type of research generally uses verbal material such as lists of words or syllables, various types of imagery like figures or pictures of common objects, and simplified versions of manual tasks, such as tracking a moving target. Studies with verbal and image material usually entail a study or exposure period followed some time later by measures of retention, such as recalling or recognizing the items studied. Manual skill learning involves repeated trials over time to assess performance improvement and occasionally tests of longer-term retention of the skill. All of these types of studies are meant to analyze learning and memory in the abstract, control for the many other variables that might influence performance, and to uncover general principles that are at work in more complex situations.

A.5.1 Basic Elements—Learning and Forgetting Curves

Learning requires time, and what is learned is not perfectly retained. Many of the earliest studies of learning and memory were concerned with characterizing the shape of learning and forgetting curves. Figure A.4 illustrates the three basic learning curve shapes that have been observed repeatedly in a variety of experiments. Curve A shows negative acceleration—rapid early performance and leveling off over time. Curve B illustrates the "sigmoid" curve—initial positive followed by negative acceleration to the asymptote. Curve C shows positive acceleration—a slow early rise with increasing gain over time. Various factors influence the shape of the learning curves, such as the difficulty and meaningfulness of the material to be learned. Such curves are characteristic of both verbal material and perceptual-motor skill learning. All show the basic importance of time—with more time, learning improves.

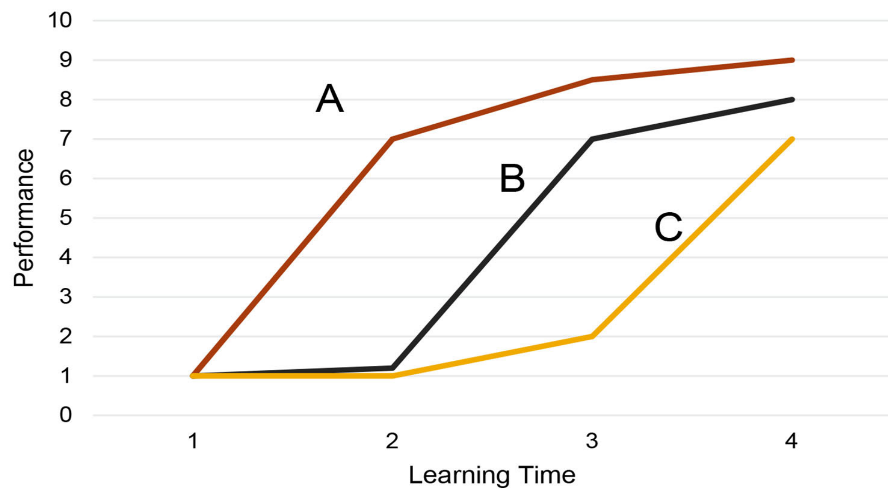

Figure A.4. Idealized learning curve shapes (McGeoch and Irion 1952) showing improved performance with increased time learning.

Foundational studies of human learning and memory were conducted by Hermann Ebbinghaus Ebbinghaus (1885/1964). The basic philosophy of Ebbinghaus' work was to simplify the problem of learning to one that could be manipulated and measured; this approach has been the hallmark of experimental psychology since that time. Ebbinghaus used himself as the single subject in these studies and employed three-letter "nonsense syllables" as learning material to control the meaningfulness of the items. His approach was to read a list of syllables of various lengths to the beat of a metronome for precise timing. Various studies were performed to determine the effect of list length, number of repetitions, and duration of the retention interval. The specific method used by Ebbinghaus to quantify memory was known as the "savings" method. This entailed measuring the number of repetitions and time utilized for initial learning and comparing those quantities with the number of repetitions and time required for relearning material to mastery (mastery was defined as a single repetition of the list without hesitation).

Of particular interest in relation to the question of initial learning and refresher training were Ebbinghaus's studies of lists that were learned to varying levels of mastery and then subsequently relearned to complete mastery on the following day. The savings, or relearning score, was based on the average amount of time required to master a list of 16 syllables—31 repetitions, or 1270 seconds. The results of this study are shown in Figure A.5, which plots the original Ebbinghaus data in terms of the percentage of "relearning time saved" for material that had been studied to varying degrees of performance. As can be seen, even for material that had been studied to nearly twice the level of learning for mastery (64 study trials), a substantial amount of relearning was required. For material that was studied to a single mastery test (32 study trials), more than half of the original learning time was required to re-master the information.

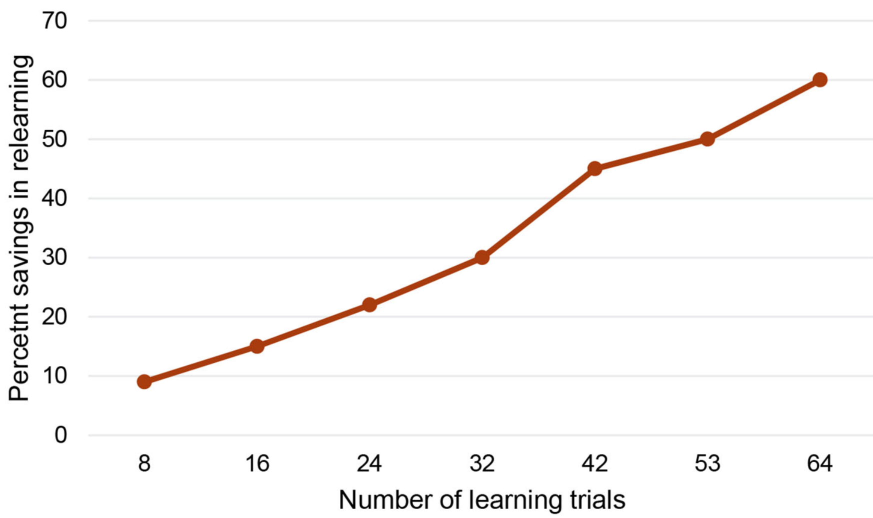

Figure A.5. Percentage saved (remembered) in relearning needed to master a list of 16 syllables 24 hours after various amounts of initial study.

An alternative way of viewing learning and memory is to measure how much information is retained over various periods of time. This question was also addressed by Ebbinghaus using the savings method of measurement. In this study, he learned lists of 13 syllables, requiring an average of 1090 seconds to master. Subsequent measurement of the relearning time required at intervals of 19 min through 31 days showed a substantial loss of the initial learning. Figure A.6 shows these data, illustrating more than 50% information loss over a day, and nearly 80% over a month.

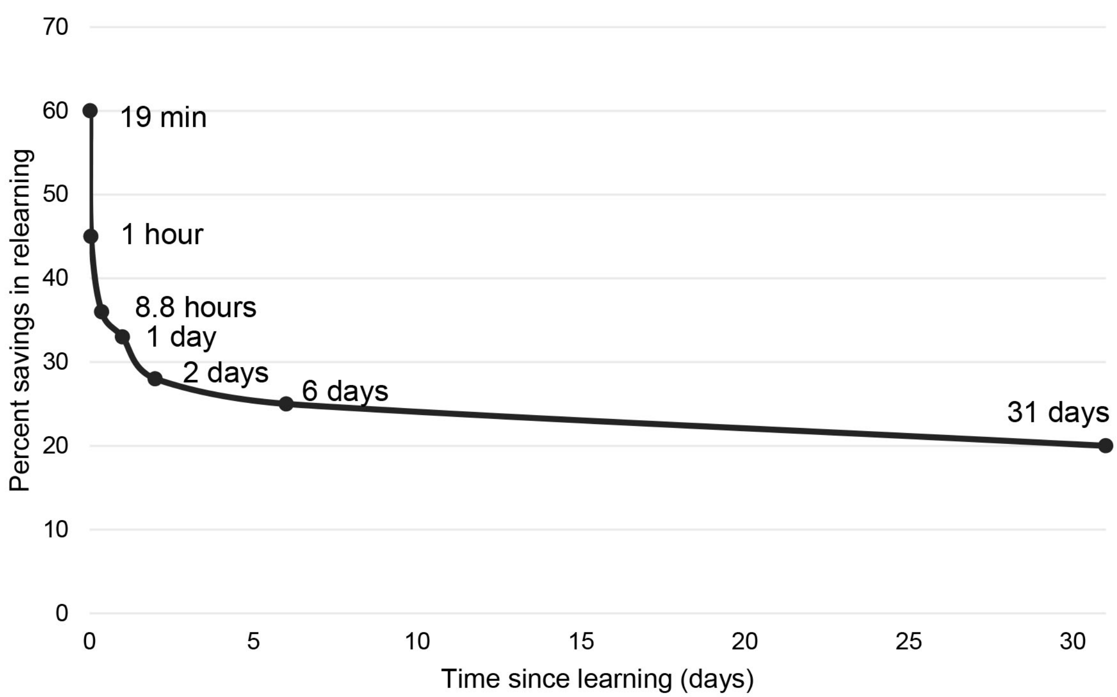

Figure A.6. Percentage saved (remembered) in relearning to master a list of 13 nonsense syllables (Ebbinghaus 1885/1964).

The form of the learning and forgetting curves illustrated above have been demonstrated across a wide range of materials, types of tasks, and durations (Stroud 1932; Kellogg 1946; McGeoch and Irion 1952; Taylor and Smith 1956; Tulving 1967). Bahrick (2000), for example, describes work with face and language memory illustrating retention levels over decades (to be discussed further below).

Substantial effort has been expended evaluating the specific form of learning and forgetting curves in attempts to fit quantitative functions (e.g., Thurstone 1919; Stevens and Savin 1962). The general consensus is that both learning and forgetting curves can be characterized as power, exponential, logarithmic or hyperbolic functions, depending on the specifics of the situation studied (Rubin and Wenzel 1996). There does not appear to be a general application of learning or forgetting curve functions to the prediction of learning time or amount of retention in a prospective manner, although a recent study of radiologists (Pusic et al. 2011) used individual learning curves and desired performance criteria to classify trainee progress. Additionally, in studies of military training, models have been developed to forecast skill retention (see volume 2 on Training and Instructional Psychology).

A.5.1.1 Relevance for NDE Training

This early work on learning and forgetting has been repeatedly replicated and illustrates two key principles: (1) more study or practice leads to greater initial learning, and (2) the amount of information retained depends on the time since learning. While it is tempting to generalize from these findings, there are many other factors that influence learning and retention, which will be discussed further below. In general, meaningful material, i.e., something with prior associations such as words, take less time to learn than completely abstract items such as nonsense syllables or random lists of digits or letters. Additionally, the greater the amount of material to be learned, the more time it takes (Hovland 1951; Baddeley 1999). Baddeley (1999) encapsulates these results in the following way:

> The process of learning shows neither diminishing returns nor the snowball effect, but obeys the simple rule that the amount learned depends on time spent learning – if you double the learning time, you double the information stored. In short, as far as learning is concerned, you get what you pay for….this is known as the total time hypothesis. (p. 72)

Baddeley also points out that while this general relationship exists, there are ways to get better value for the learning time invested. These approaches involve such methods as spacing of practice, overlearning, retrieval practice and other influences to be discussed in sections that follow.

A.5.2 Types of Memory

In everyday usage, people distinguish between what may be termed *immediate* and *long-term* memory. The former is taken to mean situations where people need to hold something in mind—like a telephone number—long enough to enter the numbers and complete the call. Long-term memory refers to any information that we can recall or recognize without referring to an external aid, such as a list of telephone numbers. Psychologists have developed these distinctions into a variety of sub-types, which are based on factors such as time, capacity, and type of information. These distinctions have utility in considering the memory requirements of the NDE job and training implications.

The information processing model of learning and cognition (Figure A.3) hypothesizes a temporary "working memory" and a permanent long-term store. These hypothetical constructs are meant to describe types and operations of memory rather than specific entities within the brain, although they do engage a variety of different neural structures (Schachter et al. 2000). The following discussion summarizes the key types of memory concepts that have relevance for work performance and for the development and maintenance of knowledge necessary to perform various NDE jobs.

A.5.2.1 Working Memory

The concept of working memory was originally based on a computer storage metaphor in which information was stored for short periods but quickly lost if not "transferred" to long-term memory (Atkinson and Shiffrin 1968). These metaphors were useful in describing specific tasks studied during that period, which addressed issues concerning memory for radio communications, telephone and postal codes. With repetition or rehearsal, such codes are well-learned and easily recalled at later points in time; however, preventing rehearsal or adding distraction reduces the ability to retain information over the short term and impedes the establishment of durable memory. Additional research with amnesic patients illustrated that short-term memory can remain intact, while acquiring new long-term memories was not possible (anterograde amnesia) (Baddeley 1992).

Working memory as described by Atkinson and Shiffrin (1968) was conceived as a means for activating and holding limited amounts of information already stored in long-term memory. Specific numbers, such as the digits 0–9, are established in memory; new combinations of those digits such as a phone number are not; thus, it is holding in mind the combination and order of new data that represents the function of working memory (see also James 1890). A great deal of research has shown that working memory is limited in capacity on the order of 3 to 4 "chunks" (Cowan 2001, 2015). A "chunk" is a single item such as a digit or a meaningful group of items such as words, phrases or categorizations of elements. Various task demands—such as interference from other sources in the environment—can degrade working memory capacity (Baddeley 2012).

The current conception of working memory represents it as a central executive mediating function between the environment and long-term memory, similar to attention. As such, working memory is a central element in complex cognitive processing, including initial learning and subsequent skilled performance (Ericsson and Kintsch 1995). Experts in a particular domain, such as medical diagnosis, show better working memory performance with regard to patient details; this is thought to be based on activation of categories or "chunks" from long-term memory that support better immediate recall. Their expertise lies in the ability to develop and use the chunks as retrieval cues for other information; on standard digit-span tasks, experts perform at the same level as non-experts.

A.5.2.2 Long-Term Memory

The concept of long-term memory was originally influenced by William James's (1890) distinction between primary and secondary memory. Primary memory—that which is currently "in mind"—is now considered by researchers to be working memory as discussed above. Secondary memory is "knowledge of a former state of mind after it has already once dropped from consciousness" (James 1890). That is, there is an implied permanence of the original knowledge. Secondary or long-term memory is presumably the process by which humans maintain and access the whole range of personal experiences, academic learning, and non-verbal memories involving other senses such as sound, touch, taste, movement, and emotional responses. Much research throughout the 20th century was concerned with how information is represented and retrieved, as well as how other material interferes with memory.

A diverse body of laboratory, neuropsychological and animal experimentation has led to the formulation of several systems of long-term memory (Figure A.7). These include non-declarative perceptual representation and procedural elements, as well as declarative memory systems, which have particular relevance in the overarching process of learning, retaining and applying knowledge in complex performance. Declarative memory entails knowledge of specific facts and events and is generally considered to comprise semantic (language-based) and episodic (time-based) elements.

Spatial memory tends to be investigated as a separate form of mental representation but can be considered as a form of blended semantic (place, features) and episodic (familiarity) systems. There are also elements of procedural memory in the practical aspects of wayfinding and navigation.

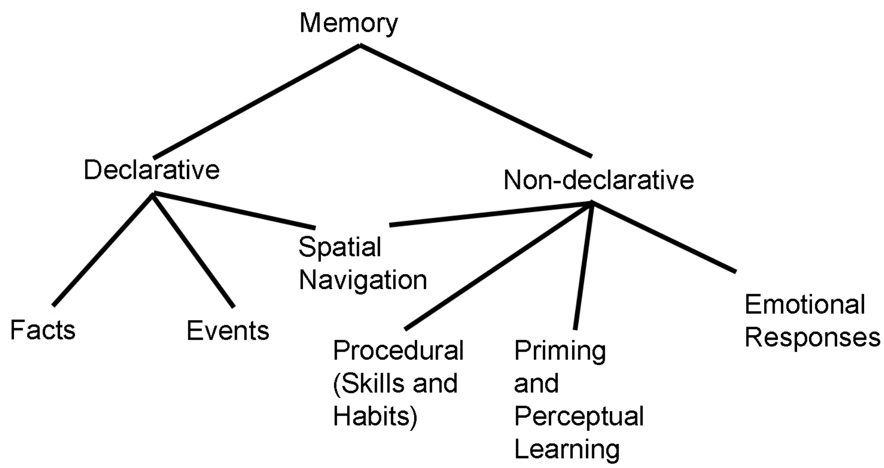

Figure A.7. Structure of long-term memory systems (Squire 2004).

A.5.2.2.1 Perceptual Representation System and Perceptual Learning

The Perceptual Representation System (PRS) is a memory construct proposed to account for a range of experimental and neuropsychological findings that suggest memory for perceptual objects—word fragments, objects, patterns, etc.—is distinct from explicit recall or recognition of those objects. The PRS is based largely on a variety of findings of *priming* effects, i.e., the facilitating effects of prior exposure to related material such as words, line drawings of objects, or faces in tasks not requiring explicit learning. After such exposure, subjects are shown reduced or degraded versions of the information and asked to name or categorize the item; this might entail several letters of words, obscured elements of figures, or abstract versions of faces previously presented. In these circumstances, prior exposure—with no intent to learn or memorize—leads to better and faster identification than similarly degraded items with no prior exposure (Tulving and Schacter 1990).

Figure A.8 is a fragmented figure in a noisy background that may require considerable time for some people to identify. Without prior exposure to related materials, identifying the object requires giving some people hints and cues (Roediger III and Srinivas 1993). Some people, for example dog owners, may immediately see the figure of a Dalmatian. This type of rapid response reflects the operation of perceptual priming.

The PRS is considered an *implicit* memory system, i.e., one which does not involve conscious recollection that material has been seen before and for which people do not make explicit attempts to recall material. Figure A.9 illustrates the type of visual material that people are asked to identify or classify following varying levels of prior exposure to similar items. Implicit memory allows people to accomplish tasks such as object recognition without having seen the exact item before.

Figure A.8. Photograph by RC James (from Roediger III and Srinivas 1993) illustrating fragmented features. The figure in the center is a Dalmatian.

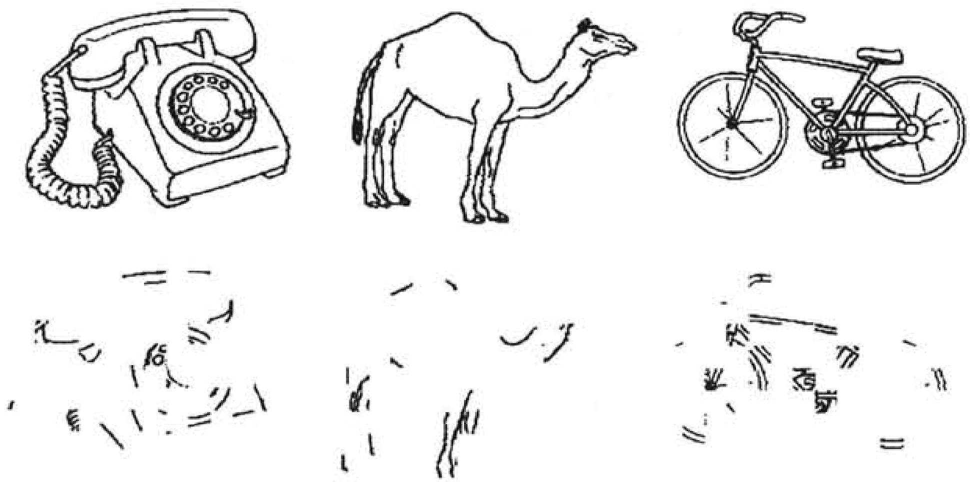

Figure A.9. Full and fragmented images used in visual priming studies (Snodgrass and Feenan 1990).

Priming corresponds to everyday experience in a variety of ways, including the ability to recognize musical pieces from a few notes or driving route wayfinding based on prior experience as a passenger without having actively navigated previously. It occurs without direct awareness. Priming has also shown to be intact in patients who otherwise have severely disrupted long-term memory (Squire 2004), indicating that aspects of material that is studied but not recallable can still influence performance. The effects of original exposure to pictures have been shown to elicit priming as long as 17 years later (Mitchell 2006).

Priming is a specific case of a more general process known as *perceptual learning*. Perceptual learning refers to “experience-induced improvements in the pickup of information” (Kellman and Massey 2013). Numerous laboratory tasks have shown that with experience, subjects are able to more quickly discover information that is relevant to their task domain, and also become more fluent, i.e., they can extract that information with greater speed and ease. This might be considered a form of learning to “see” with greater proficieincy—as expertise develops, the ability to discern patterns in complex visual stimuli improves over novices who cannot discern such patterns. A variety of application-oriented studies suggests that relatively simple instruction on image interpretation, followed by classification of large numbers of images, can lead to novices performing nearly as well as experts for radiographs, histopathology slides, and electrocardiograms (Kellman 2013).

A.5.2.2.2 Procedural Memory

Memory for sequential operations that people are able to execute without conscious recollection of individual steps or elements is considered *procedural*. This type of memory is distinguished from others in that it develops over a period of practice, and the specific perceptual-motor actions occur without people necessarily being able to verbalize exactly what they are doing. Procedural memory is considered “knowing how” as distinct from “knowing that.” Everyday examples include knowing the proper form of language grammar without being able to explain the rules and the sequential motions of physical sports such as bicycling or swimming without overt awareness of the individual movements. Knowlton et al. (2017) also distinguish between goal-oriented procedural memory—that which is used to accomplish a discrete task—versus habit memory, which entails established responses in specific situations that are not necessarily appropriate to the circumstances. An example is making a usual left turn on a route normally travelled to work but unneeded because you are going elsewhere.

Psychologists have demonstrated dissociation between implicit procedural learning and memory in amnesic and normal subjects in a variety of tasks. These include learning to read mirror image text, learning to draw reversed images, various types of tracking, and higher-order activities such as process control. The standard finding in these activities is that subjects improve their performance over time but when queried about details of the task are unable to verbalize them.

Procedural learning takes place in cognitive activities as well as perceptual motor tasks. For example, Berry and Broadbent (1984) had subjects learn to control the output of a simulated industrial process based on labor force adjustments or the interpersonal style of a simulated co-worker based on the friendliness of interactions. Multiple trials led to improved performance on each simulation, but subjects scored at chance level on questions regarding what they did to achieve the results. Thus, their knowledge of "how" clearly improved, although their overt knowledge of "what" did not. Similar dissociations have been shown between prior verbal instruction and task performance, i.e., improvements in performance were no different without instructions than with them.

Procedural learning and application are generally thought to proceed through a series of stages (Fitts 1964; Rasmussen 1983). Initial skill acquisition is relatively deliberate and slow, with increasing association between specific inputs and outputs that finally reaches an autonomous state of direct linkage between sensory inputs and appropriate responses. This progression has been formalized in various models, with that of Rasmussen (1983) having the most applied flavor (Figure A.10). Initial formal, knowledge-based behavior leads to decisions about what needs to be done, planning how to do it, recalling rules for application, and finally taking action. As learning progresses, appropriate responding becomes more automated, eventually leading to much less conscious deliberation and more immediate output. Progressive "automating" of task sequences is a feature in numerous cognitive theories.

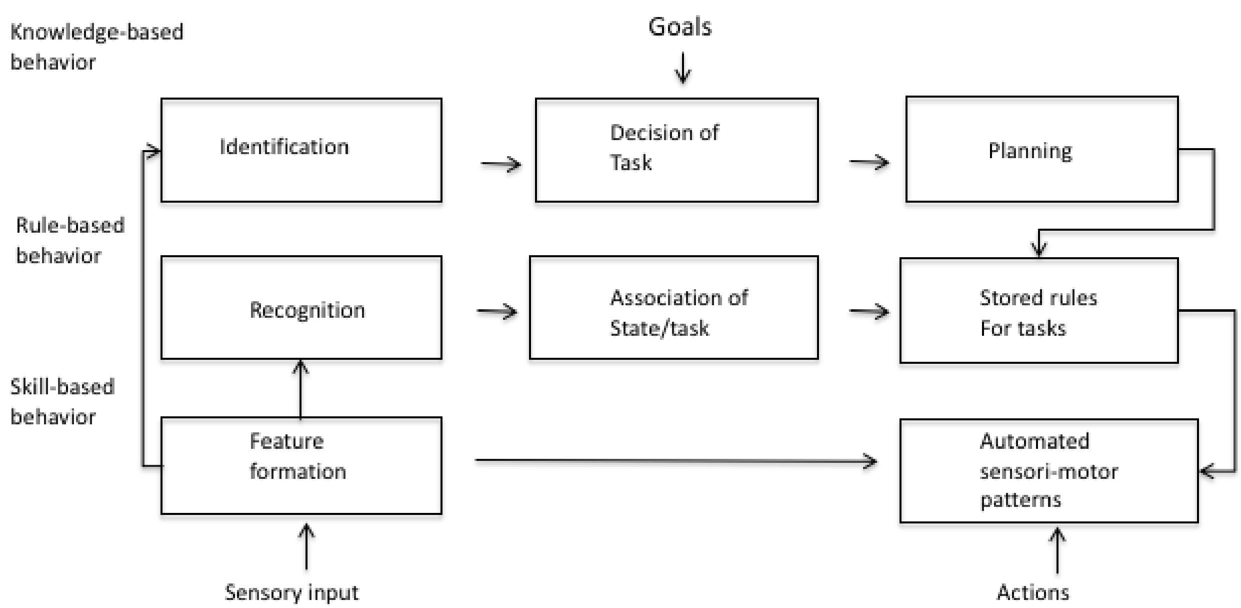

Figure A.10. Knowledge-rule-skill hierarchy (Rasmussen 1983).

A.5.2.2.3 Declarative Memory

As described above, declarative memory is considered “knowing what” as distinct from procedural “knowing how.” It involves conscious recollection of facts, order relationships and is the basis for how traditional educational curricula are structured. Further distinctions have been made to define semantic and episodic memory, the former being language-based and reflecting our knowledge of the world (e.g., a bicycle has two wheels), while the latter is time-based and reflects personal knowledge of when events occurred, such as whether you have ever seen a bicycle before. The two types of memory operate together in everyday life, but the distinctions are useful for considering how different types of memory develop, change over time, and are applied in various settings (Tulving 1972).

Studies of semantic memory have generally been concerned with performance, such as judging whether specific items are members of a category (e.g., robin-bird) and a variety of other types of related judgments (Shoben 1992). Theories of semantic memory are concerned with the structure of knowledge representation and is closely aligned with artificial intelligence. Of primary concern for this review is the notion that knowledge is represented as a structure of related units, nodes, clusters and categories, and that various relationships between these units can either speed or slow decision making. For example, people are quicker to judge that a robin is a bird than they are that a chicken is a bird; this is the *familiarity effect*. It has been suggested that activation of nodes in the semantic memory structure can spread and thereby facilitate speed of decisions and also induce interference (Collins and Loftus 1975). More recently, a variety of connectionist models have been described (Thomas and McClelland 2008), with the goal of enhancing machine learning. Figure A.11 illustrates a conceptual structure of a small semantic network.

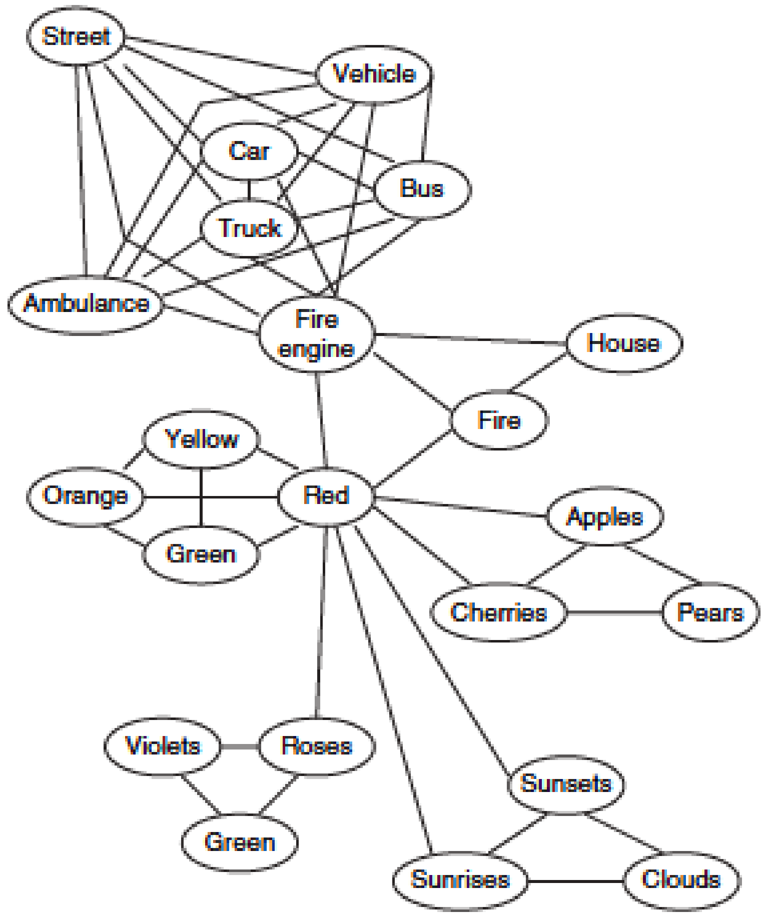

Figure A.11. Illustration of a semantic network (Collins and Loftus 1975).

Episodic memory is self-referential, i.e., we make judgments about events and material on the basis of whether we have experienced them before. In a standard list-learning type of study, the “episodes” are the individual items on the list, and performance is a function of the complexity of the material, length of the list, degree of attention, etc. A robust finding in episodic list-learning studies is that the “level of processing” of the items on the list leads to variation in retention. For example, if people are required to judge whether two words are in the same font type, their recall of those words will be inferior to a condition in which they judge whether the words have similar meanings (Craik and Lockhart 1972; Craik and Tulving 1975). More recent studies of autobiographical memory (episodic by definition) have shown that younger people tend to recall more temporal event detail (when something happened), while older adults tend to report facts and figures (what happened and where) (Levine et al. 2002).

The reality of declarative memory is that semantic and episodic systems work together in everyday life. There are clear differences in modes of operation, type of material remembered, susceptibility to alteration, etc. Table A.1 lists some of the principal distinctions (Tulving 1984).

Table A.1. Selected characteristics of episodic and semantic memory (adapted from Tulving 1984).

Episodic	Semantic
Encodes events, occurrences	Encodes facts, ideas, concepts
Organized temporally	Organized conceptually/hierarchically
Memory more context-dependent	Memory less context-dependent
Experience of remembering	Experience of knowing
Illustrates general education material	Structures general education material

A.5.2.2.4 Spatial Memory and Navigation

The analysis of spatial memory and navigation is somewhat divergent from the traditional, verbal learning procedures used to study episodic and semantic memory. However, people clearly develop long-standing memories of spatial layouts and are adept at navigating routes that were once not known, often without the need to consciously recollect specific waypoints. Thus, spatial memory in application might best be considered a blend of declarative (semantic and episodic) and procedural systems. The notion of "cognitive maps" was first suggested by Tolman (1948) on the basis of animal studies and is now a commonly used term.

In a recent review of the topic, McNamara (2017) characterized the aspects of spatial memory and knowledge to include objects or places (e.g., a hill or a park), route knowledge (the paths between places), environmental shape such as elevation and orientation of pathways, and survey knowledge (the overall configuration of an environment in a common reference system). Such knowledge tends to be organized hierarchically, such as cities within states in relation to compass directions; this can lead to errors, such as judging Reno to be east of San Diego (Stevens and Coupe 1978).

A diverse body of research suggests that spatial memory and navigation skills are learned incrementally; landmarks and their sequential order are identified fairly quickly, while metric (distance) information is developed more slowly. Over time, people develop overarching survey knowledge that can be useful in devising alternate routes. Chrastil (2013) has delineated the component cognitive processes involved in establishing spatial memory as the basis of navigation:

- Recognition of places
- Learning sequences
- Identifying decision points
- Learning appropriate responses
- Forming associations between places and actions
- Goal identification
- Path integration

A variety of human and animal research supports these general processes, which contribute to the integrated activity of knowing and finding one's way.

A.5.3 Learning versus Performance

Learning was defined above as a change in behavior resulting from experience. The behavioral change is measured by some type of change in performance on tasks or tests. In the traditional classroom environment, test performance is the most widely used measure, which is often criticized as not necessarily reflecting what a person knows. This latter issue points to a distinction between learning and performance. For learning experiences to be most useful, the knowledge and skills developed need to be durable and flexible, i.e., “accessible across periods of disuse… and in the various contexts in which they are relevant” (Soderstrom and Bjork 2015).

There is abundant empirical evidence that learning and performance changes are often distinct, i.e., learning can occur in the absence of evident behavioral and performance changes, suggesting learning may not be enduring. This section discusses the research addressing the learning-performance distinction. It is noteworthy that NDE training and qualification are oriented toward performance demonstration, and prior research suggests that junior examiner experience hours are often aimed at “checking the box” in order to meet the requirement to sit for a performance demonstration examination (Sanquist et al. 2018), rather than providing the most beneficial learning environment. Thus, an understanding of the basic research in this area can facilitate training and evaluation processes. The material in this section is based on the recent review of learning and performance distinctions by Soderstrom and Bjork (2015).

A.5.3.1 Latent Learning

Learning that occurs without obvious changes in behavior is referred to as “latent.” There are many examples of latent or passive learning in daily life, such as learning a route to work while a passenger on a bus—an activity that does not require active engagement or incentive. It is not evident that people are actually learning in this situation; however, when required to do so most people are able to execute the route if driving themselves (the incentive being getting to work).

The original research demonstrating latent learning was performed in the early 20th century. An early set of issues in the psychology of learning at that time was the extent to which reinforcement (reward) was necessary—could animals learn in the absence of reward? This question was addressed by Blodgett (1929) and Tolman and Honzik (1930). In these experiments, three groups of rats were put in a maze under different conditions of reinforcement: none, reinforcement after reaching the goal, and delayed reinforcement. The results of this work are shown in Figure A.12. Animals in each group were taken out of the maze when they reached the goal box. Day 11 shows that the group that received food reinforcement upon reaching the goal exhibited a consistent reduction in errors over the 17-day test period. The non-reinforced group showed a minor reduction, whereas the group that received a food reward on Day 11 showed an immediate drop in error rate. From that point on, the performance of this group was equivalent with the always-rewarded group.

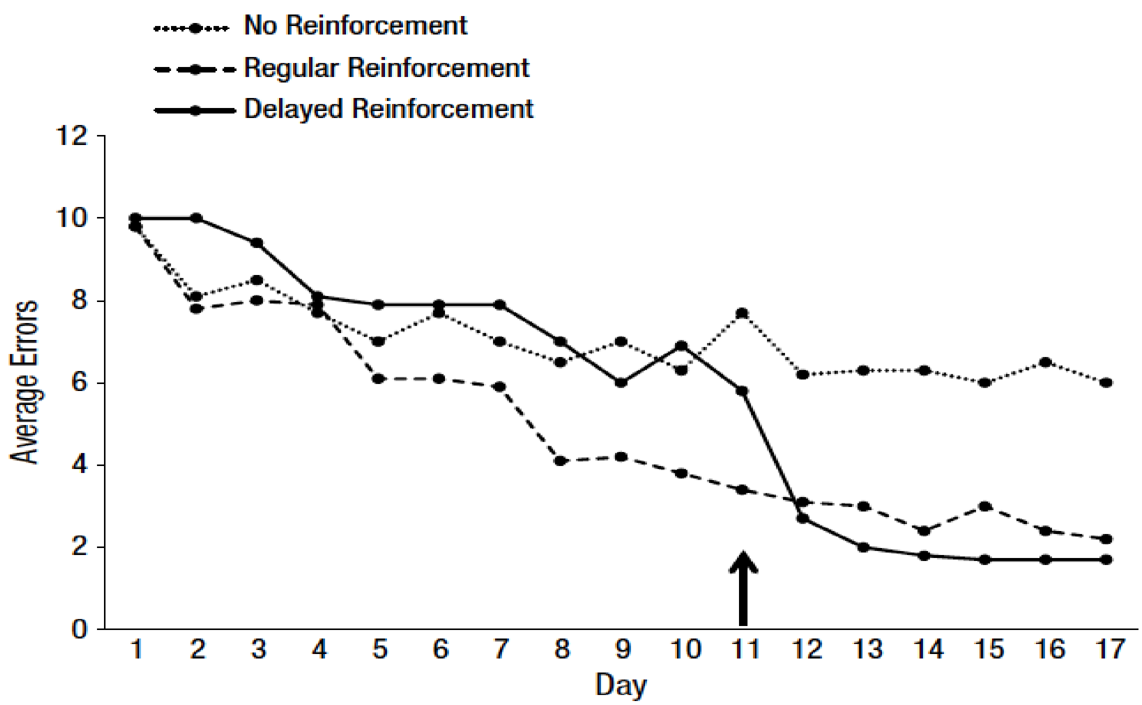

Figure A.12. Average number of errors committed while searching for the goal box. Arrow at day 11 shows when food reward was introduced for the delayed reinforcement group. Adapted from Tolman and Honzik (1930).

The results were interpreted to mean that learning *did* take place; it was just not evident in performance until a reward was provided. Tolman and others developed this method over a considerable period of time, demonstrating numerous variations on the basic phenomenon (Tolman 1948), which led to the introduction of the term “cognitive map.”

Latent learning has also been demonstrated in humans (Stevenson 1954). In this experiment, children explored objects in search of a key that would open a box. The environment also had non-key objects. During later tests, the children performed substantially better than chance in finding objects that were irrelevant to the main task, suggesting that they had acquired information about their location. Stevenson also points out that latent learning in humans had generally been studied in somewhat different experimental setups, i.e., not maze learning, under the label “incidental learning,” which is discussed in the next section.

Latent learning is an oft-cited example in the basic psychological literature. The research was originally performed in the context of evaluating whether reinforcement was a necessary condition for learning to occur. Changes in research and theoretical orientation over time led to a reduction and eventual cessation in latent learning work. However, the phenomenon was robust and suggests that exploratory learning occurs in the absence of external motivation and reward.

A.5.3.2 Incidental Learning

Incidental learning is defined as “learning which apparently takes place without a specific motive or a specific formal instruction” (McGeoch and Irion 1952). These authors further point out that

> Much of the learning of everyday life is of this incidental sort. One drives along a road, walks down a street, meets friends and talks to them, and afterwards remembers much that seems to have been entirely separated from any specific motive which had been operating or any reinforcement which may have occurred. (p. 210)

Research on incidental learning is similar to the studies of implicit (nondeclarative) memory discussed above. An orienting task is provided to subjects that does not involve instructions that they will later be tested for what they have been exposed to. Subsequent testing reveals that information was retained, despite no requirement to learn, at a higher level than for material not included in the orienting task.

A prototypical example of incidental learning is provided by an experiment reported by Eagle and Leiter (1964). In this study, three different groups of subjects studied material with three different instruction conditions: (1) study for a later test of recall, i.e., reproduce as many words as possible from memory when asked, (2) judge whether each word was a noun, verb or adjective, and (3) remember the words for a test and perform the judgment. The group required to learn and remember recalled more words than the other two groups; however, a recognition test (have you seen this before?) showed that the incidental task (Condition 2–semantic judgment) led to better retention on a recognition test. The judgement/remember task (Condition 3) yielded intermediate results. This study not only shows the learning effect of incidental instructions, but also the difference between retention tests—recognition is superior for the incidental group (Condition 2), while recall is superior for the intentional group (Condition 3). This suggests different cognitive strategies were employed in the two tasks and foreshadows later work on depth of processing and elaborative encoding in learning.

Incidental learning is a topic and experimental procedure that is presently being used to study attentional and perceptual processes. The basic question of interest is whether people develop the tendency to bias their attention in a particular direction based on undisclosed statistical regularities in displays that they inspect for the presence or absence of a target. These studies have demonstrated that during a visual target search, people become more familiar with background distractors whether they are repeated in the same or different locations, as shown by faster reaction time (Hout and Goldinger 2010). Further, there is evidence that people learn whether certain locations are more likely to have targets in complex displays; they also have a tendency to move their eyes to those locations first (Jiang et al. 2014).

The general pattern of results from incidental learning studies over many years is that regardless of instructions or disclosure of statistical properties of information, people do learn without specific intent. The nature of the learning depends upon the specific situation at hand but appears to operate at the level of both verbal/semantic information, pattern recognition, and the direction of attention in space.

A.5.3.3 Relevance for NDE Training

The distinction between learning and performance is directly applicable to concerns about NDE training. It is clear from the results discussed in this section that learning does occur in the absence of demonstrated performance (latent learning) and that people acquire and retain information on an incidental basis when performing another task. These findings suggest that in training NDE examiners, there is an impact of the experience obtained by junior personnel as they acquire experience hours. This can include developing a “cognitive map” of complex work environments, heuristics for how to scan various structures, and mental models of procedure execution. Specific knowledge and performance tests for this type of latent and incidental learning do not exist for NDE field experience, but interviews with examiners suggest its importance ("the first couple of years were a blur… it takes time to know where you are going") (Sanquist et al. 2018).

A.5.4 Overlearning

Overlearning is another area in which learning and performance can be dissociated. Tests of learning show the level of mastery of material; conditions during the learning phase of experiments are analyzed to determine impacts of variables such as material content, organization, amount of study time, and repetition upon material mastery and retention. The research on latent and incidental learning suggests that performance may not indicate the degree to which learning has occurred. Similarly, a test of mastery or retention may not tell the whole story—if continuous or periodic study of material occurs after the first test of mastery, does further learning occur and would this have an impact on subsequent retention? Put another way, just because a student has passed a test is not an indication of long-term retention or skill development. Conversely, will further study increase retention? This is the domain of *overlearning,* and a number of studies suggest a positive impact, at least for short-term retention.

Ebbinghaus's original work on nonsense syllable learning involved conditions of overlearning. For some lists, he studied more than twice the number of trials required for first error-free reproduction. On tests of relearning 24 hours later, the overlearned material required approximately 30% fewer trials to regain mastery of the lists. Krueger (1929) examined overlearning with longer retention intervals (up to 28 days) and found that overlearning did lead to greater retention over time and required less relearning, but the effects appeared to grow smaller over time. Improvements in methods to control for interference from prior learning were applied by Postman (1962), who showed that overlearning tended to have the greatest impact for material that is most difficult to retain, i.e., the middle items in a list.

Throughout the subsequent period, overlearning was used as an experimental procedure to test the impact upon changing response requirements to a stimulus situation, primarily in animal studies. The concern was the extent to which transfer of training was facilitated or impeded by overlearned response patterns. The general finding from studies during this period is that overlearning facilitated transfer of training to similar response patterns and impeded learning of highly different responses. An everyday example is the tendency to look to the left when crossing the street as a pedestrian; travel to countries where left-sided driving is the norm can be confusing and dangerous for pedestrians. In some highly traveled tourist areas, "look right" is painted on the curb at crosswalks. This example can be considered a "habitual" response in terms used by Knowlton et al. (2017) to describe procedural learning, i.e., an overlearned response that is not applicable to the situation that evokes it.

During the 1960s–1990s, the few human studies of overlearning that were performed used a very diverse range of tasks and retention intervals, generally reporting a positive effect, but only for a relatively short period (e.g., 1, 4, or 7 days predominantly and very few over 14 days). Because of the generally positive findings, overlearning took on the characteristic of traditional lore as far as effectiveness. Researchers during this period concluded that overlearning was an effective technique for retention of knowledge and skills, despite the early evidence that the effects diminished over time.

In 1992, (Driskell et al.) conducted a meta-analytic review of studies with human subjects. This analysis compared the effect sizes of various degrees of overlearning for physical and cognitive tasks and assessed the impact of retention interval. The overall finding was that for cognitive tasks, overlearning produced positive effects that diminished to zero at 38 days post-learning. This was not the case for physical tasks, such as machine gun assembly, which showed an impact of overlearning at 8 weeks post-training. Driskell et al. (1992) suggested that the physical tasks may have been rehearsed, consciously or otherwise; however, the loss of overlearning benefits for most tasks over time led them to suggest that refresher training be given on the basis of learning performance "half-life," i.e., the point at which original learning declines to 50% or less of the original level.

Rohrer and colleagues (2005; 2006; 2009) have evaluated overlearning within the context of massed versus spaced practice. Overlearning typically involves massed practice in which material is studied to a criterion of mastery and is then studied further in the same session. Retention is measured later by various means and in some cases constitutes additional learning, so there is a potential confounding effect. Spaced practice entails studying material to a specified level of performance and delaying further study to some later point in time.

In a series of studies designed to specifically assess the effects of overlearning at short and extended retention intervals, Rohrer et al. (2005) compared the performance of near-mastery performance with complete mastery for lists of relatively unknown city-country pairs to be learned. The "lo-learn" condition entailed 5 trials, which led to 85% average correct performance, while the "hi-learn" condition entailed 20 trials, leading to 97% correct performance. In the hi-learn condition, 58 of 63 subjects produced at least 3 correct recitations of the city-country pairs, whereas 44 of 67 lo-learn subjects produced only one perfect trial. Subsequent memory testing at 1, 3, and 9 weeks showed superior performance for the hi-learn condition at 1 week, which diminished substantially at the two longer retention intervals (Figure A.13).

Using similar procedures, this result was replicated in a second experiment that showed the effects of overlearning to be reduced to zero at 4 weeks (Rohrer et al. 2005).

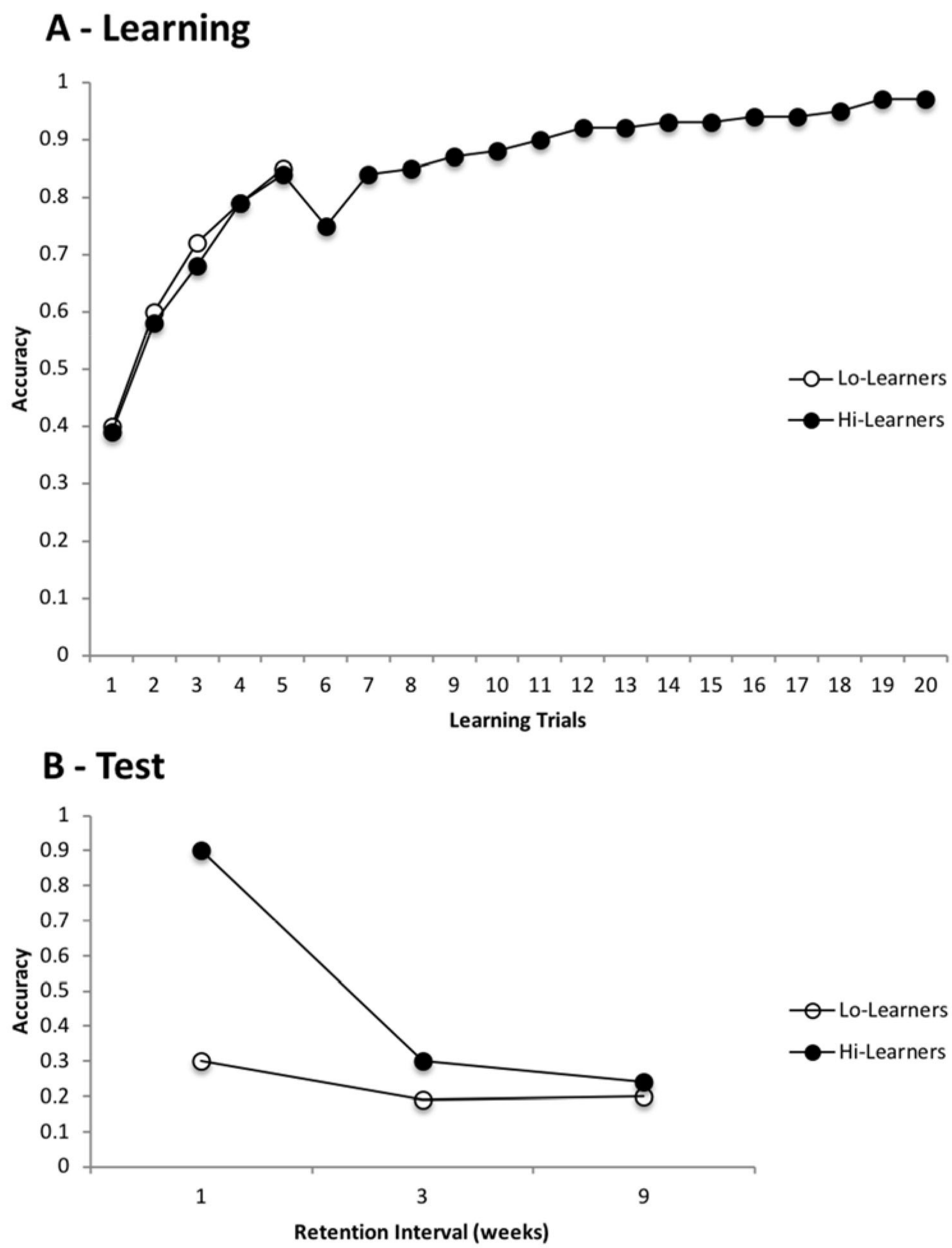

Figure A.13. Learning and subsequent test performance for sub-mastery (lo-learner) and overlearning (hi-learner) conditions (Rohrer et al. 2005).

Subsequent work by this group (Rohrer and Taylor 2006; Rohrer 2009) analyzed the impact of concentrated study time (massed practice), with the same amount of study time spread over separate sessions. Rohrer (2009) had subjects learn lists of little-known author-story pairs in the following conditions: high masser (20 massed trials), low masser (10 massed trials), high spacers (20 spaced trials), and low spacers (10 spaced trials). Massers completed all their study trials within a single study session; spacers divided their study trials across two sessions separated by two weeks. Massers were tested two weeks after their learning session, and spacers were tested two weeks following their last learning session. The 20-trial massed spacing study condition was associated with overlearning, i.e., on the last 10 of the trials, most tests of memory were perfect. The 20-trial spaced condition subjects achieved mastery at the end of the first session (10 trials), but their performance dropped to 36% on the first learning trial of the 2nd week session. Therefore, these subjects spent most of the remaining trials in the spaced condition relearning, meaning they had not overlearned.

The paradoxical results are shown in Figure A.14—the subjects who performed at near mastery during the first session but did not overlearn (high spacers) showed substantially better recall in the final test session.

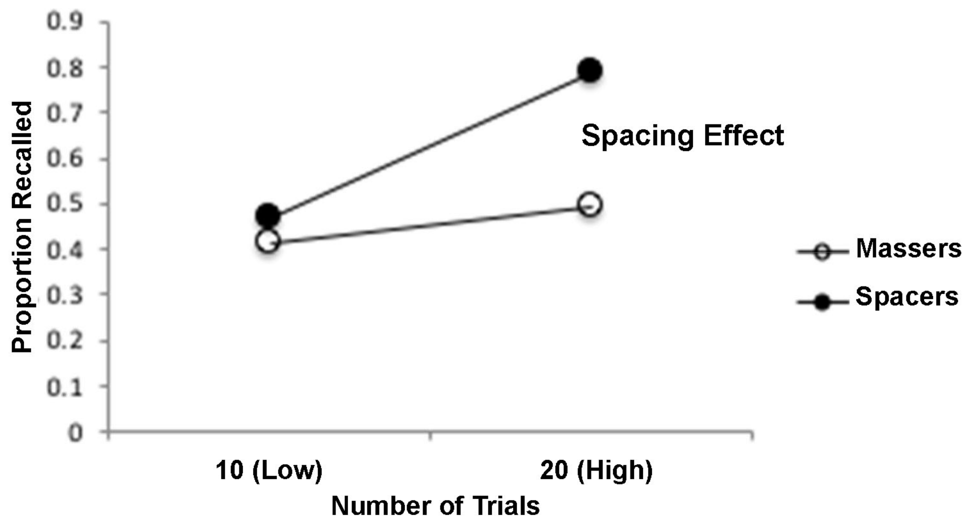

Figure A.14. Memory performance for overlearning (massers) and distributed learning (spacers) based on Rohrer (2009).

This type of effect, i.e., the beneficial impact of spacing study time out rather than continuing to study in a single session (the basic definition of overlearning) leads to superior long-term retention. This has been shown for mathematics problem solving as well as list learning (Rohrer and Taylor 2006). These researchers point out that much of traditional mathematics and other educational curriculum learning is based on concentrated study of one type of problem or knowledge, which would appear to be counterproductive for long-term retention.

A.5.4.1 Relevance for NDE Training

Overlearning—study of material beyond the point of mastery in close succession to the original study period—is a common characteristic of practical education and training. The fundamental technical principles of NDE tend to be taught in contexts that emphasize material mastery before progression to the next stage. The practice requirement of ASME Code Appendix VII simply defines a quantitative requirement, such as 8 hours annually, with no recommendation for how that time is obtained; it might be all done in one day to meet the requirement, a common practice with continuing education, or obtained as availability of practice samples permit.

Research indicates that overlearning is a transitory phenomenon, at least as measured in constrained laboratory tests. While it may lead to better test performance in the short term, the benefits diminish quickly over time. The research of Rohrer and colleagues (2009) suggests that partitioning the study time spent in overlearning across separate sessions can enhance retention. This is the more general phenomenon known as the “spacing effect,” which is discussed in detail in the next section.

It should be noted that overlearning is also used as a term to describe the impact of advanced levels of training (Bahrick et al. 2013). This entails a progressive curriculum of material that moves from basic to more advanced and uses the basic material as the foundation for increasingly complex concepts and operations, a typical characteristic of academic disciplines. In this case, overlearning of the basic material is achieved through a periodic repetition in the context of a wider variety of learning material and has been shown to enhance retention over many years. This will be discussed further in the section on long-term maintenance of knowledge.

A.5.4.2 Practice Distribution and Variability

Repetition is a universal aspect of learning. Across all organisms, repeating the association between two events strengthens its representation in memory. It has also been known since the earliest studies of human learning and memory that "with any considerable number of repetitions a suitable distribution of them over a space of time is decidedly more advantageous than the massing of them at a single time" (Ebbinghaus 1885/1964, p. 89). The spacing of repetitions in learning is a very general effect and can be seen in non-human species as well (Gerbier and Toppino 2015).

Research on the spacing effect has gone through periods of exploratory and parametric study in an attempt to characterize the nature of the phenomenon with a variety of methods (McGeoch and Irion 1952; Melton 1970) and to develop explanatory theories (see Cepeda et al. 2006; Carpenter 2017). There has been a recent amplification of interest in the spacing effect in view of its potential application to educational settings. The linkage to the fundamental issue of refresher training frequency is obvious—is there an optimum interval for practice to enhance long-term retention of knowledge? This section discusses the nature of the spacing effect and illustrative research studies.

A.5.4.3 Spacing Effects

A demonstration of the spacing effect is illustrated in a verbal learning experiment by Madigan (1969). In this study, subjects viewed lists of 48 words, among which were some items that were repeated twice and at intervals ("lags") of 0, 2, 4, 8, 10, and 32 intervening words. After viewing the words, presented one at a time, subjects recalled as many as they could, regardless of order. Figure A.15 shows that increasing the space between repetitions improves the ability to recall items later.

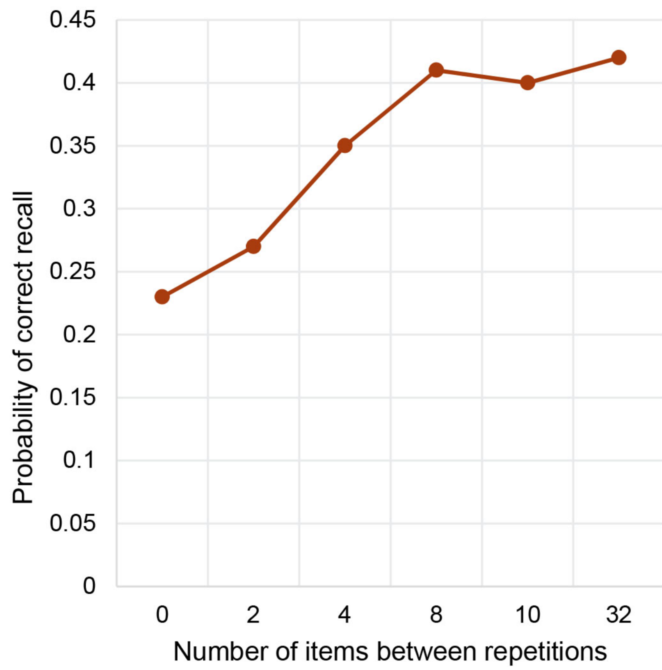

Figure A.15. Probability of correct recall for items repeated with intervening material between repetitions (adapted from Madigan 1969).

A large number of research studies have addressed various aspects of the spacing effect to try and understand why it occurs and what the optimal conditions are for achieving the effect. The review by Cepeda et al. (2006) documents 254 studies using nearly 15,000 experimental participants. The benefits of spaced versus massed practice apply across diverse types of learning material, including motor skills, verbal, mathematical, and imagery.

Analyzing why the spacing effect works is important for understanding how it might be used in educational and training settings. Carpenter (2017) reviews the various theories of the beneficial impact of spacing repetitions in learning. The principal theories include:

- Deficient processing—mass repetitions result in less attention than spaced material
- Encoding variability—spaced repetitions occur in multiple psychological and physical contexts, thus yielding more cues for remembering
- Study-phase retrieval—repetition reminds people of earlier occurrences, and the earlier presentations are retrieved (a “rehearsal” of the material)
- Consolidation—spaced repetitions permit time-dependent neural processes to occur, which are disrupted with massed repetitions.

A variety of evidence can be marshalled for these various theories, which are not mutually exclusive. Recent work by Bahrick and Hall (2005) suggests that spaced repetitions and testing are important in generating feedback from *retrieval failures*, which then served to re-focus learner attention on material that is not as strongly encoded in memory.

Use of the spacing effect in applied situations is demonstrated by a study conducted by Baddeley and Longman (1978) who investigated how to optimize postal workers’ typing ability for new postal codes. Should they learn the system as quickly as possible—suggesting a massed presentation approach—or would their ability be enhanced over the long term with more distributed learning? The results indicated that less study time per day (1 hour) over a longer period led to faster and more durable learning than a 4 hour per day schedule compressed into a shorter period. The distributed practice group expressed frustration at the seeming inefficiency of their schedule, which is a common perception that leads people into counterproductive massed-practice study strategies. This perception of inefficiency can reinforce the tendency to use massed practice (e.g., re-reading) as a study strategy, which has been shown to be relatively ineffective (Dunlosky et al. 2013). When learners are required to allocate their study strategies between massed and spaced repetitions, they appear to select material they judge to be most difficult for massed practice (Pyc and Dunlosky 2010), suggesting that individual judgments of learning are important moderators.

A practical question that occurs in many settings concerns optimal practice spacing in relation to when knowledge will be tested. The basic issue is the durability of learned material, which depends not only on when and how it was learned, but also on when attempts are made to retrieve that information from memory. The specific issue of refresher training frequency is directly related to this question.

Most of the research on the spacing effect has shown a benefit of distributed practice regardless of the retention interval—seconds, days, months or years (Cepeda et al. 2006). There are sometimes exceptions to this general finding, however, indicating that closer spacing between repetitions may be more advantageous for short test intervals. The preponderance of research on spacing effects and retention have used test intervals of less than one day, raising questions about the generality of the findings for more durable learning and retentions.

The most detailed study of the relationship between the spacing of learning sessions and the time interval of testing is reported by Cepeda et al. (2008). This study used internet-based learning and testing to evaluate 26 combinations of spacing gap and retention interval; spacing gaps varied from 0 to 105 days, and the retention interval varied from 7 to 350 days. Subjects learned a set of obscure facts (e.g., inventor of snow golf [Rudyard Kipling]) to mastery in the first learning session and, following the gap, received one additional presentation of the material. After the study sessions, memory was tested at the various intervals by both recall and recognition methods. The results showed that the optimal spacing of learning trials is dependent upon the length of the retention interval; this relationship is illustrated in Figure 2, which shows that shorter spacing gaps are best for shorter test delays, and moderate spacing gaps for longer test delays. Across all conditions, the optimal gap produced a 64% increase in recall and a 23% increase in recognition.

For the longer retention intervals (70 and 350 days), 21 days between learning presentations produced the best recall. For the 7- and 35-day test delays, study spacing of 1 and 11 days yielded the best results. The non-linear relationship between study spacing and test delay suggest that it is not possible to precisely define the optimal gap between learning events to enhance long-term retention. From a practical standpoint, however, review mid-way through a semester course curriculum is likely to produce better exam performance than immediate or end-of-week post-class review. Similarly, for retention over a longer period, e.g., several years, review after several months post-learning may potentially double the amount remembered (Cepeda et al. 2008).

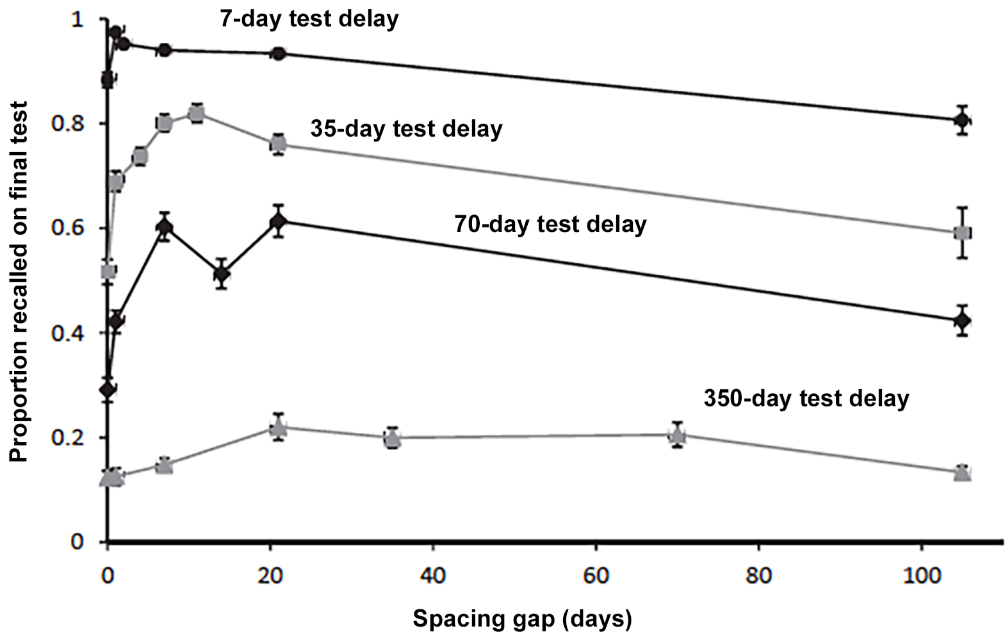

Figure 2. Recall of facts for combinations of learning spacing gaps and retention intervals (based on Cepeda et al. 2008).

A.5.4.4 Variation of Practice Conditions and Material

Studies of varying practice conditions and material are conceptually related to the spacing effect but tend to focus more on how practice is executed and the specific material practiced. An example comes from an experiment reported by Kerr and Booth (1978) in a study of motor skill learning in children. The skill to be learned was tossing a beanbag onto a target at a distance of 3 feet. One group practiced exclusively at the 3-foot distance, while another group practiced at distances of 2 and 4 feet. Skill testing occurred at the 3-foot distance, and the variable practice group was more accurate than the single-distance practice group. A number of similar studies of motor skill learning under conditions of variable practice have shown similar results (discussed in Soderstrom and Bjork 2015).

In the domain of verbal and conceptual learning, similar findings have been obtained in experiments that used *interleaved* study material, i.e., the same general type of information but different content or format. For example, Taylor and Rohrer (2010) controlled the temporal spacing between items and showed that students were better able to develop geometric problem-solving skills when they practiced on different types of problems in a session rather than a single type of problem in a blocked practice session. Similarly, Kang and Pashler (2012) showed that the ability to discriminate between paintings by various artists was facilitated by simultaneously showing examples of the painting styles rather than by blocked presentation of different paintings by the same artist.

A number of other studies related to types of information to be learned have been conducted to try and determine the generality of the interleaving effect (see Carpenter 2017 for review). The findings for motor skill learning tend to be more clear-cut, i.e., variable conditions of practice lead to enhanced skill and generalization across situations. Informational material (pictures, problem solving, concepts) tend to show more complex effects, and the beneficial impact of interleaving seems to depend on the degree of distinction between the general category of material and specific features that distinguish examples of the category (Carpenter 2017).

A.5.4.5 Educational Applications

Despite a long history of research demonstrating the beneficial effects of study spacing in laboratory studies, it is only relatively recently that specific investigations in classroom educational settings have been conducted. Carpenter (2017) provides a detailed review of research that has employed school children, university students, and medical students, using material as simple as spelling words and as complex as histopathology diagnostic skills. In all cases, spaced study sessions resulted in superior performance on later tests of knowledge. The specific findings of these studies will be described in the next chapter, where the focus is on training and instructional research.

A.5.4.6 Relevance for NDE Training

Distribution and variation in practice is a set of concepts and research findings with broad general applicability in education and training. NDE training has many characteristics of standard classroom education, as well as more specific, skill-oriented practice. Practical application of distributed practice can be challenging for instructors who are already working with a fixed curriculum and have limited time for instruction, and less time for modifying existing schedules of learning. Further, students routinely judge that massed practice leads to better learning, whereas performance shows the opposite. This may be most immediately concerning in the area of practice training required for NDE examiners. If the eight required hours are done all in one sitting, e.g., with practice samples, there will be much less retention of knowledge than if practice were performed in two-hour intervals four times throughout the year. Similarly, alternating different types of samples and welds would capitalize on the beneficial effects of practice variation.

A.5.5 Retrieval-Based Learning

A.5.5.1 Background

The research discussed thus far has focused on variables associated with the study phase of learning, such as how long material is presented, how many times and over what types of spacing intervals study occurs, prior presentation of related material, etc. However, as indicated by some of the results in the previous sections, there is more to learning and retaining information than simply studying it. As Bahrick and Hall (2005) suggested, the role of memory retrieval failure may serve as a cue that induces additional effort or attention during relearning or additional attempts to remember. More generally, the study-phase retrieval hypothesis of the spacing effect suggests that longer spacing intervals induce memory retrievals, whereas massed spacing entails immediate memory. Additionally, in everyday life and work settings, what we need to remember is usually not studied prior to each time we recall that information, and it is often the case that we recall more the harder we try or with the passage of time.

The facilitating effect of repeated attempts to remember material learned previously has been known for many years (Roediger III and Butler 2011). Early studies in classroom settings demonstrated that recitation improved retention more than additional study (Gates 1917) and that testing without feedback a week after learning enhanced performance on a final test given later (Spitzer 1939). Interest in this effect waned as psychologists focused their work on characterizing the nature of forgetting, but interest was re-kindled by infrequent but consistent reports in the literature regarding improved memory performance associated with testing.

Given the central role of testing in NDE certification and qualification and the lack of a requirement for testing with annual refresher training, this topic warrants attention. In this section we will discuss the findings from a variety of studies conducted between the late 1960s and the current time which suggest that the use of strategic testing can positively impact the retention of knowledge, independent of additional study.

A.5.5.2 Hypermnesia

It is a common experience for people to be unable to recall something at a particular time—a name, place, word, etc.—but for it to seemingly spontaneously occur to them sometime later. A very familiar example of this is the “tip-of-the-tongue” experience, in which we feel as if we know what we cannot recall, but the memory later recurs without conscious effort (Brown 1991).

The recovery or growth of seemingly forgotten material is in direct contrast with the early work on forgetting reported by (Ebbinghaus 1885/1964) and many others more recently. However, the basic phenomenon of improved memory over time without additional learning has been known for over 100 years (Payne 1987). Theoretical shifts reduced attention to what was termed “reminiscence” for the middle part of the 20th century. There was a growth of interest in this effect in the 1970s as research in a variety of non-experimental psychology domains seemed to indicate the ability to recover “lost memories,” which had clear applied ramifications in areas such as eye witness testimony reliability, self-judgments of learning, clinical neuropsychology, and advertising. The term hypermnesia was proposed by Erdelyi and Becker (1974) to describe recovery of previously learned material as distinct from amnesia, which describes failure to recall what has been previously learned.

Improved recall of studied material over time without relearning was reliably demonstrated in a series of studies by Erdelyi and Becker (1974), Erdelyi and Kleinbard (1978), Roediger and Thorpe (1978), and Roediger and Payne (1982). The basic findings suggest that hypermnesia, i.e., recovery of previously unrecalled material, is a function of a number of previous attempts to recall. The earlier studies suggested that time was the critical variable—longer delays led to increased recovery of previously unrecalled items. Roediger and Payne (1982) examined the interaction between test delay and number of prior retrieval tests and showed that if tests were delayed, there was no improvement in memory over more immediate tests. However, multiple retrieval tests, regardless of delay, led to increased memory performance over successive tests (Figure A.16).

The results of hypermnesia studies may seem surprising in view of the previously discussed forgetting functions originally demonstrated by Ebbinghaus (1885/1964). Substantial differences existed, however, between the Ebbinghaus procedures and more contemporary research, including his original learning to mastery, use of nonsense syllables, and measures of savings in relearning. The contemporary hypermnesia studies tend to use single presentations of word lists or pictures and measure proportion of correct recall. Thus, these procedural differences are the likely source of disparity. More generally, forgetting does occur, as shown by the imperfect recall in the hypermnesia studies. The important point of this work is that memory does remain and that information can be recalled over time and multiple retrieval attempts. The importance of multiple retrieval tests is a key element in the hypermnesia effect and will be discussed further below.

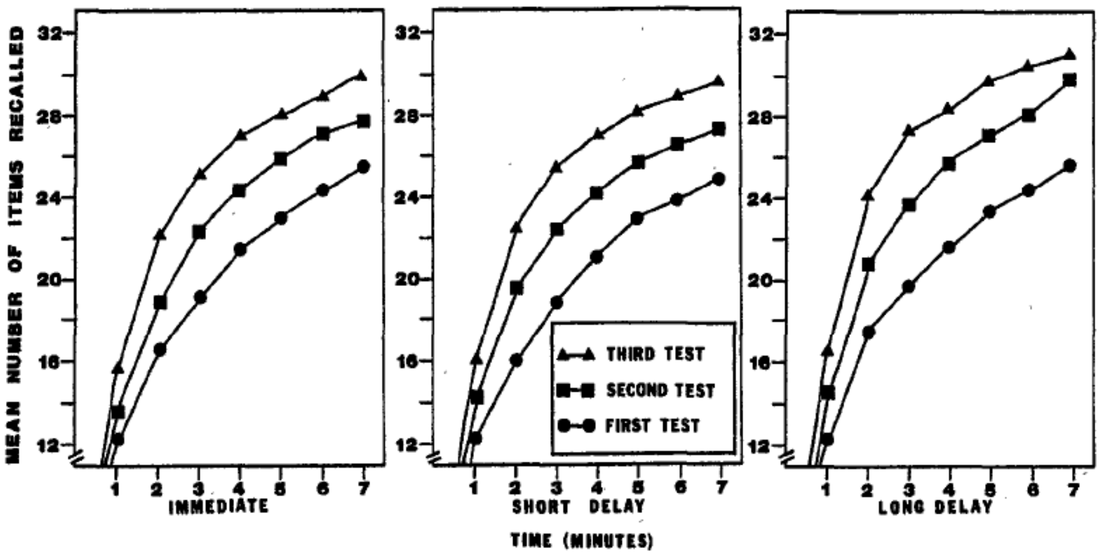

Figure A.16. Cumulative recall for multiple tests at immediate, short and long delays (Roediger and Payne 1982).

A.5.5.3 The Testing Effect

The results of the Roediger and Payne (1982) study on hypermnesia suggest that there is an effect of testing memory that is independent of the passage of time, i.e., more tests yield better performance. Testing can have two types of influence on memory—mediated and direct (Karpicke 2017). Mediated effects include encouraging continuous study and performance feedback, which can refocus study efforts on specific material. Direct effects of testing are a result of taking the test itself, without feedback or further study. Mediated effects engage additional learning processes, whereas direct effects seem to entail a strengthening of the memory (see Karpicke 2017 for a discussion of the multiple theories of the testing effect). A large number of experiments have shown that several memory tests can have as much beneficial impact as a single learning study session (Rowland 2014).

A demonstration of the direct effect of testing is shown in an experiment by Roediger and Karpicke (2006). In this experiment, subjects studied passages of text about scientific subjects. In one condition, they studied the material twice (study-study), and in another they studied once and then took a recall test (study-test). At the end of the experimental session following a 5-minute delay, another test was given. Subjects were subsequently tested 2 days and 1 week later. The results, shown in Figure A.17, clearly indicate that while multiple study sessions slightly enhance performance on immediate tests, the study-test combination yields better long-term performance. Research addressing the optimal schedule for retrieval testing has been performed by various groups (reviewed by Karpicke 2017), but no firm conclusions can be made regarding the optimal approach, e.g., equal intervals or expanding intervals of retrieval practice; the benefits are likely to be a combination of type of material and vulnerability to forgetting.

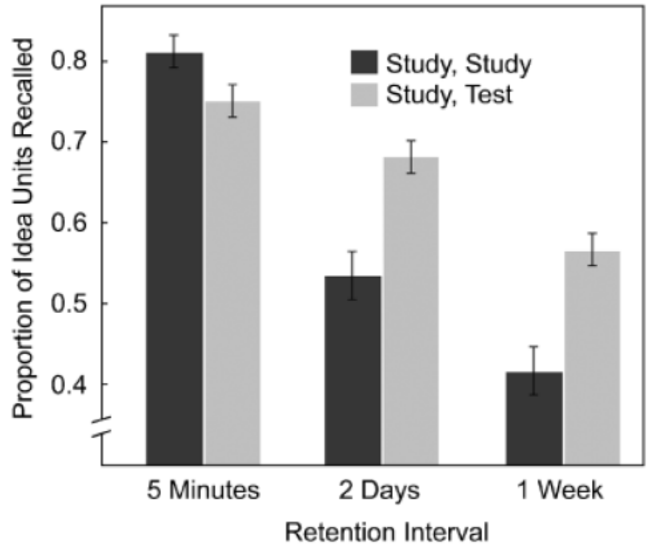

Figure A.17. Average proportion of idea units recalled from test passages after three retention intervals (Roediger and Karpicke 2006).

The testing effect has been studied primarily in college students but has also been demonstrated in school-age children and older adults (55–65). Additionally, abstract symbols and spatial positions on maps show the benefit of retrieval testing.

A technique related to retrieval practice is known as *elaborative interrogation*. This technique involves exposing learners to factual material while questions are asked regarding "why" some action was taken or event occurred (e.g., "the hungry man got in the car"). This type of condition is compared to those involving passive reading or providing an explanation of the material ("the hungry man got in the car to go to a restaurant"). When subjects are required to make their own explanation of the material, overall recall of material (e.g., "who got in the car?") is substantially better than the other conditions (Dunlosky et al. 2013). The prevailing explanation of this effect is that prompting "why" reasoning leads to better knowledge integration and retrieval cues. Similar effects occur with a technique called *self-explanation*, in which learners explain their thinking as they solve problems; using this technique concurrently or retrospectively appears to enhance memory and transfer (Dunlosky et al. 2013).

There are potential negative consequences of testing, primarily involving exposure to incorrect alternatives on multiple-choice tests. Experiments have shown that incorrect alternatives from multiple-choice testing can intrude in later tests of a different kind (e.g., short-answer) (Marsh et al. 2009). However, providing answer feedback following multiple-choice testing can reduce intrusions (Karpicke 2017). Feedback will be discussed more thoroughly in a separate section below.

A.5.5.4 Educational Applications

A substantial amount of research has been conducted in the past 20 years to evaluate implementation and impacts of retrieval testing in classroom settings (Karpicke 2017, Table 2). This work will be discussed more extensively in the chapter concerning training and instructional psychology, but several results deserve highlighting at this point. The use of clicker systems, where instructors can pose questions "on-the-fly" and have students answer by means of a handheld device, have shown benefits on subsequent test performance (feedback is provided in these systems). Paper and pencil methods can also be used and show similar effects. There is evidence that for students who experience test anxiety, repeated testing reduces their nervousness about exams.

A.5.5.5 Relevance for NDE Training

The use of testing in NDE examiner certification occurs at multiple levels—initially to show competence on required curriculum material and later to demonstrate performance for specific procedures and types of welds. While practice is required annually, testing is not. The research discussed in this section supports the use of testing, with or without feedback. Thus, the periodic re-examination required of NDE examiners—which generally does not provide individual sample feedback—may be thought to induce a testing effect, particularly since there is some evidence that re-tests after initial failure yield better performance (Cumblidge and D’Agostino 2016). The lack of a testing requirement for the annual 8 hours of practice does not necessarily mean that individual organizations or inspectors do not test; it does mean that we have little knowledge about what is actually done. The research on retrieval-based testing suggests that spatial and symbolic information benefits from testing, so it may be appropriate to address how testing with complex patterns from ultrasonic exams might benefit from this effect.

A.5.6 Feedback

A.5.6.1 Background

Studies of the testing effect discussed above can be categorized into those that provide feedback and those that do not. A meta-analytic review by Rowland (2014) shows that memory retrieval tests with feedback enhance performance much more than those that do not. Early conceptions of feedback consider it as a specific instance of reinforcement, as shown in the studies of animal learning by Thorndike (1911). Thorndike’s work involved placing a cat in a puzzle box which had a loop of string connected to the door. Over time, the cat learned that pulling the string with its paw would open the door and allow access to food. The basic result is summarized in the Law of Effect, as follows:

> Of the several responses to be made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation so that when it recurs, they will be more likely to recur.

The Law of Effect was further elaborated to suggest that behaviors associated with discomfort would be less likely to recur in the future. These findings and theories formed the basis for what was later termed reinforcement, i.e., a consequence of behavior that increases its likelihood. Early learning theorists studied many variants of reward, punishment, and time schedules in an attempt to understand precise relationships.

The term “feedback” took on more formal usage in human psychology in the 1950s and 1960s on the basis of servo-control theory in engineering (Adams 1968). An excellent definition of feedback is provided by Weiner (1954):

> Feedback is the control of a system by reinserting into the system the results of its performance. If these results are merely used as numerical data for criticism of the system and its regulation, we have the simple feedback of the control engineer. If, however, the information which proceeds backwards from the

> performance is able to change the general method and pattern of the performance, we have a process which may very well be called learning.

In distinction to behavioral stimulus-response theories, which were essentially feed-forward systems, Miller et al. (1960) proposed that the results of behavior served an error-correction function in a closed loop model. In this theoretical formulation, feedback serves to adjust behavior to achieve the desired goal and is distinguished from "reinforcement" in behaviorist terms, which simply serves to strengthen a response emitted by an organism. This distinction is very important for subsequent developments in cognitive psychology since it provides the basis for an internal or "mental model" developed through learning that can be modified on the basis of feedback.

Studies of feedback in motor learning and visual tracking have identified some of the key variables. These include whether feedback is intrinsic to the task, such as proprioceptive input from the muscles, or extrinsic, such as a visual indicator showing tracking deviation. Additional considerations include whether feedback is immediate or delayed and whether it is specific regarding error or more general, as in "correct-incorrect." It is generally found that disrupting intrinsic feedback impedes learning and that providing extrinsic feedback specific to performance improves learning. Delay and greater precision of feedback have been found to enhance motor learning (Salmoni et al. 1984).

A.5.6.2 Feedback in Cognitive Tasks

Feedback in cognitive tasks, such as list or concept learning, or responses on test items in analogues of classroom testing, such as multiple-choice and short-answer, has been the subject of much study and debate. Early studies of "teaching machines" and programmed instruction were conducted in the theoretical framework of reinforcement, i.e., feedback is meant to increase the probability of the correct response and thus should be delivered immediately. Despite substantial variations in type of learning material, type of feedback, and delays, an overall pattern from decades of research has emerged favoring somewhat delayed feedback, specific to the accuracy of a student's response (Kulhavy 1977). This section illustrates these principles with several recent experimental reports.

Pashler et al. (2005) studied the differences between providing right/wrong feedback versus correct answer feedback in a verbal learning study. This study entailed learning Ugandan-English language pairs, and evaluated three levels of feedback: none, right/wrong, and providing the correct answer. Tests of retention were given on Day 1 (original learning), Day 2, and one week later. The results showed that correct answer feedback yielded better performance on the Day 2 and one-week tests. Further, this effect was specific to incorrect responses in the original learning test, i.e., correct answer feedback had no facilitating effect if the response was already correct.

A more detailed investigation of type of feedback was conducted by Fazio et al. (2010), using sentences from which subjects learned specific facts upon which they were later tested. An additional manipulation of this study examined the effects of whether material was reviewed after testing, and the impact of answer feedback vs. right/wrong feedback. This study provides a more comparable approach to how people actually learn in educational settings, i.e., receiving a form of feedback during the learning process and an opportunity to review prior to final testing.

The results of Fazio et al. (2010) showed the superior effect of correct answer feedback upon final test results. They also found that for test items upon which errors were made, right/wrong feedback increased test performance when students were not provided an opportunity to review material. This result suggests learning is enhanced even when correct answer feedback is not provided, particularly when there is no opportunity to review the material again.

A detailed examination of feedback delay was reported by Butler and Roediger (2008) in conjunction with the nature of multiple-choice tests. In these studies, subjects read 400-word passages containing critical facts, which were then tested via multiple-choice questions. During the test, correct answer feedback was provided either immediately after the question was answered or at the end of the series of 42 test questions (ranging from 7–14 minutes, depending on the presentation order of questions). One week later, subjects were given a test of final recall to assess retention over a longer period. Assessment of the effects of number of alternatives on the multiple-choice questions was also made using questions with 2, 4, or 6 alternatives.

The results of this study showed that across various study conditions, there was a graded effect of feedback, with delayed feedback yielding the highest retention (Figure A.18).

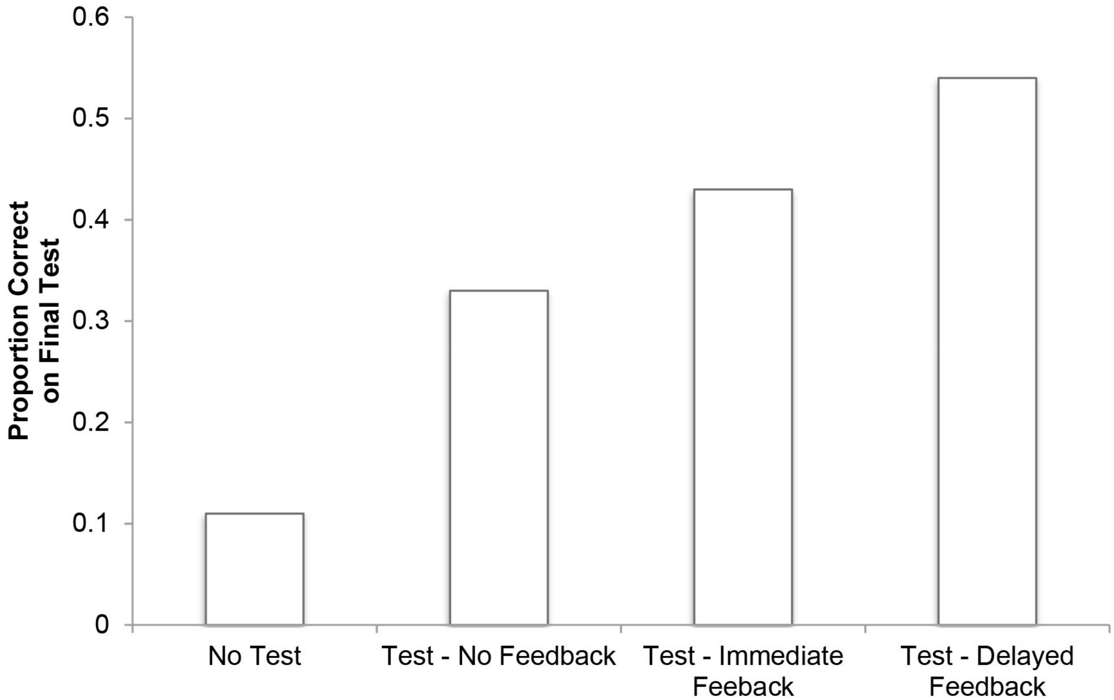

Figure A.18. Proportion of correct responses on final recall test as a function of feedback [based on Butler and Roediger (2007); cited in Roediger III and Butler (2011)].

Butler and Roediger (2007) also observed that the number of alternatives on the multiple-choice tests was correlated with intrusion errors in final recall but that if initial errors were corrected by feedback (either immediate or delayed), the effect of test format was reduced.

The laboratory studies discussed in this section illustrate the impact of types feedback and their interaction with test format and timing. Feedback of any type, whether immediate or delayed, is better than no feedback at all. Further, if it is only possible to provide right/wrong feedback due to testing logistics, this will still have a beneficial effect upon error correction, although not as strong as correct answer feedback. Delaying feedback to at least the end of the test items rather than providing immediate item-by-item feedback enhances longer-term retention. Researchers suggest that this is due to less cognitive competition from the incorrect response in working memory and also to a spacing effect—delaying correct answer feedback essentially comprises a spaced presentation of the material to be learned.

Feedback serves to reduce the intrusion of incorrect alternatives in subsequent tests; this is an important consideration, since students also learn incorrect information in the form of multiple-choice lures simply by being exposed to them. Finally, confidence in answers made on tests varies—correct answers can be made with high- or low-confidence. Evaluation of feedback provided to subjects who rate the confidence of their responses shows (1) that low-confidence, correct responses tend to be reinforced and better retained than similar responses without feedback, and (2) that errors made with high confidence are corrected with enhanced attention to the feedback provided (Butterfield and Metcalfe 2001; Butler et al. 2008).

A.5.6.3 Learning from Errors

The foregoing discussion indicates the importance of feedback for correcting errors and enhancing confidence in correct responses when learners may be uncertain. In traditional learning theory, reinforcement (feedback) is intended to shape behavior toward the desired goal and away from non-goal-directed behavior. This latter type of behavior can be considered “error” in a more traditional testing or classroom environment, i.e., the student makes the wrong response. There is a strong tradition and philosophy in education that errors are undesirable and that instruction should be oriented toward error-free learning (Metcalfe 2017). However, evidence suggests that errors provide a learning opportunity and can specifically enhance subsequent learning if appropriate feedback is provided. This section discusses the potentiating effects errors can have on learning and the use of various levels of informative feedback.

Conventional wisdom suggests that if students make an error on a test, no learning has occurred and that insufficient study time was devoted to the material. This assumption is challenged by research conducted by Kornell et al. (2009). In these studies, subjects were given the task of either generating a weakly associated word with a cue (e.g., factory – plant) or simply studying the pair. Time of exposure to the material was equated so that overall viewing time was the same. The condition in which subjects generated the associated words can be considered as a test without prior study. When subjects made an error, i.e., generated the wrong associated word (e.g., factory – building), they were provided feedback with the correct answer. On subsequent memory tests, it was found that memory was better for the words that were provided as feedback to correct erroneous responses. The results suggest that simply increasing passive study time is less effective than actively generating test responses, even if the responses are initially wrong.

Kornell et al. (2009) theorized that generating an error had an activating effect in a semantic memory network, which was facilitated when the correct response feedback was provided. This effect has been replicated a number of times and appears to occur when the errors are related in some way to the correct response, as shown by semantic analysis (Huelser and Metcalfe 2012). Huelser and Metcalfe also found that subjects were unaware of the better memory performance for error-corrected feedback; they thought that more reading time was better, indicating a dissociation between self-perceived learning and performance.

Correct answer feedback is generally more informative than right/wrong feedback, except under very specific conditions, such as the lack of opportunity to review material. Other methods of feedback include answer-until-correct, in which multiple alternatives are provided until the correct response is selected, and scaffolded feedback, in which increasingly informative cues are given to stimulate the learner to generate the correct response. Butler et al. (2008) tested answer-until-correct in comparison to correct answer feedback and found no difference in subsequent memory performance. Finn and Metcalfe (2010) tested answer-until-correct, standard correct answer, and scaffolded feedback, and found that the scaffolded approach yielded the best longer-term memory performance (1 or 2 days post-learning). Scaffolded feedback basically involved presenting increasing numbers of letters in the correct response in word-pair learning, leading the participants to generate the correct answer.

A.5.6.4 Educational Applications

The role of feedback in education has a long history with many debates and controversies. The most salient of these was discussed above—the view that reinforcement should be delivered immediately and that the goal of reinforcement is to promote "correct" performance. The laboratory studies reviewed above show that such reductionistic thinking is incorrect and that there are many subtleties involved.

Perhaps the most visible use of feedback in traditional educational settings is in the form of programmed learning, or "teaching machines." These devices date to the late 1920s (Pressey 1926), with revived interest in the 1950s (Skinner 1954). However, as discussed by McKeachie (1974), the initial promise of these machines failed to deliver, particularly as the various complexities of real-world learning intervened.

Metcalfe and Kornell (2007) reported on their research involving translation of various cognitive science principles such as answer generation, error-based learning, and feedback in a traditional classroom setting with sixth graders. Their results showed that studying vocabulary by means of computer-based presentation using cognitive principles was highly superior to conventional methods of vocabulary study such as flashcards, colored pens and other similar study aids. However, Metcalfe and Kornell found that small differences in item presentation, such as a pause between word definition and the correct response, tended to eliminate some of the beneficial effects of self-generated responses. These findings indicate the importance of careful attention to instructional design variables when adapting laboratory findings to educational settings.

Feedback in education also focuses on aspects of learning that are more general than specific factual material. This includes motivation, means of providing feedback, the role of praise, and information related to the learner's task, process, self-regulation and self-perception (Hattie et al. 2017). This research is within the domain of instructional psychology and will be more fully discussed in the next chapter.

A.5.6.5 Relevance for NDE Training

Beyond the general applicability to all forms of learning, feedback has particular relevance to the issue of recurrent training with practice samples and for performance demonstration tests. With regard to practice samples, the psychological research suggests that testing can yield benefits but only if structured in such a way as to actually challenge the trainee and provide informative feedback. We do not know at this point how individual NDE examiners engage with practice samples, but to be effective, the practice should be equivalent to inspecting a component that has not been seen before, with truth data available and provided after practice.

A consistent issue in our field interviews with NDE examiners is their concern with the lack of feedback during periodic performance testing for various procedure qualifications. They are simply told whether or not they passed. Senior NDE examiners have failed these examinations, as well as more entry-level personnel, and they have little information available to them to understand what they did wrong and to correct their errors. There is concern in the testing organization that feedback will compromise the test items, which are limited in number and expensive to produce. This understandable concern might be mitigated somewhat with a scaffolded feedback approach employed during the test examinations, e.g., providing cues without giving away the answer. Alternatively, post-test debriefing using recordings from the examination instrument might represent a method of providing feedback beyond pass/no-pass, while maintaining integrity of specific test samples. Similarly, use of a low- or medium-fidelity simulator could allow generation of an unlimited number of weld flaw variations, with concomitant performance feedback.

A.5.7 Long-Term Retention and Maintenance of Knowledge

A.5.7.1 Background

Most of the research discussed thus far has addressed learning and memory measured in laboratory situations in which the retention interval is relatively short—often 1 hour, less often over weeks, and rarely longer. Work cited above addressed spacing effects with a retention interval of 1 year. Priming in the perceptual representation system has been shown to exhibit small effects for 17 years. The early studies of Ebbinghaus suggested that as much as 70% of learned material is lost within 6 days. Subsequent studies of verbal learning (Davis and Moore 1935; cited in Hovland 1951) illustrated that while the *shape* of the forgetting curve is similar for nonsense syllables and meaningful (e.g., prose) material, the *levels* of retention are different (Figure A.19).

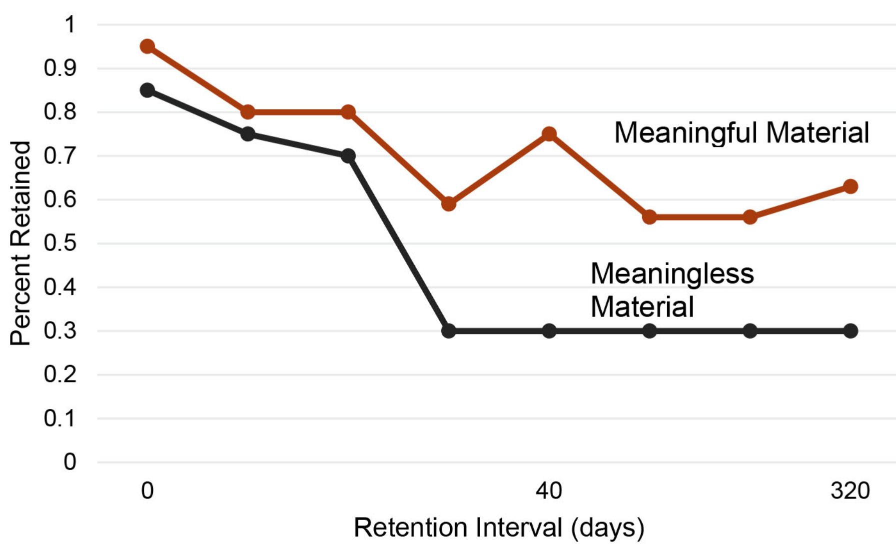

Figure A.19. Retention curves for meaningful and meaningless material (from Davis and Moore 1935; cited in Hovland 1951).

Meaningful material is retained at a much greater percentage than meaningless items and varies with how retention is measured. Meaning can also be imposed during the learning process through instructions to pay attention to particular aspects of the material such as semantic characteristics (Craik and Lockhart 1972). The method of testing memory can show varying levels of retention, such that recognition yields higher levels of retention than recall.

Further, even within a single study session, learning represents the cumulative effects of acquisition and loss of information (Melton 1963), indicating that “pure” measures of what is learned versus what is forgotten are not possible. Thus, the oft-cited myth of physicians “remembering almost nothing” from their basic science education is not borne out by a variety of different studies to be discussed below. What is at issue is how much is retained over what period of time, and how can that knowledge be maintained through periodic practice or review instruction? This section reviews research that addresses memory over the very long-term—decades and longer—and studies that focus on enhancing memory durability.

A.5.7.2 Alternative Research Methods

Adequately characterizing very long-term retention is a challenge for traditional laboratory experimental methods which need to be completed within a period of time that is practical for the researchers and the subjects. Longitudinal research follows the same group of subjects for an extended period of time, usually well-beyond the longest retention interval used in experimental studies. This type of research can involve experimental manipulations at the beginning of the study and throughout the length of the research period; subjects can serve as their own controls, so the effects of individual differences are reduced. Unfortunately, the loss rate of subjects is high for longitudinal research, and adequate final sample sizes require more participants than can generally be enrolled and retained for lengthy periods.

Cross-sectional research entails using different groups of subjects who have experienced the same basic learning material but at different points in time. For example, students who have studied Spanish in college experience the same general curriculum and can be selected for how long ago they were in school. Groups are established for whom learning is very recent, up to 50 years prior, and intervals between. Tests of knowledge are given to each group, and a variety of other data, such as frequency of use of that knowledge, original level of training (beginner to advanced), grades received in classes, etc., is collected. Multiple regression is used to control for differences in level of original learning, grades received, and a variety of other independent variables. Bahrick et al. (2013) have used the cross-sectional research approach to study retention of various types of knowledge over periods of up to 50 years.

A.5.7.3 Very Long-Term Retention Results

The Bahrick et al. (2013) publication documents a series of research studies first published in 1975 and continuing over a 40-year period that used many different types of learning material. This section selectively reviews this work and several similar studies from other research groups.

In the first reported study using the cross-sectional method, Bahrick et al. (1975) studied memory for names and faces of students in the same high school graduating class. In this study, 392 participants were assigned to one of nine groups having graduated from high school between 2 months and 57 years prior to being tested. Knowledge tests were constructed from names and pictures in high school yearbooks, and participants were tested on a variety of different measures, including name recall, name recognition, face recognition, face-name matching, and picture cuing to retrieve the associated name. The results indicated that name and picture recognition and the ability to match names to pictures were retained at high levels—above 80%—for 25 years. This level of performance dropped somewhat beyond 35 years but was still above 60%. In contrast, the ability to recall specific names or to associate names with pictures declined steadily over time. The results of this study are shown in Figure A.20.

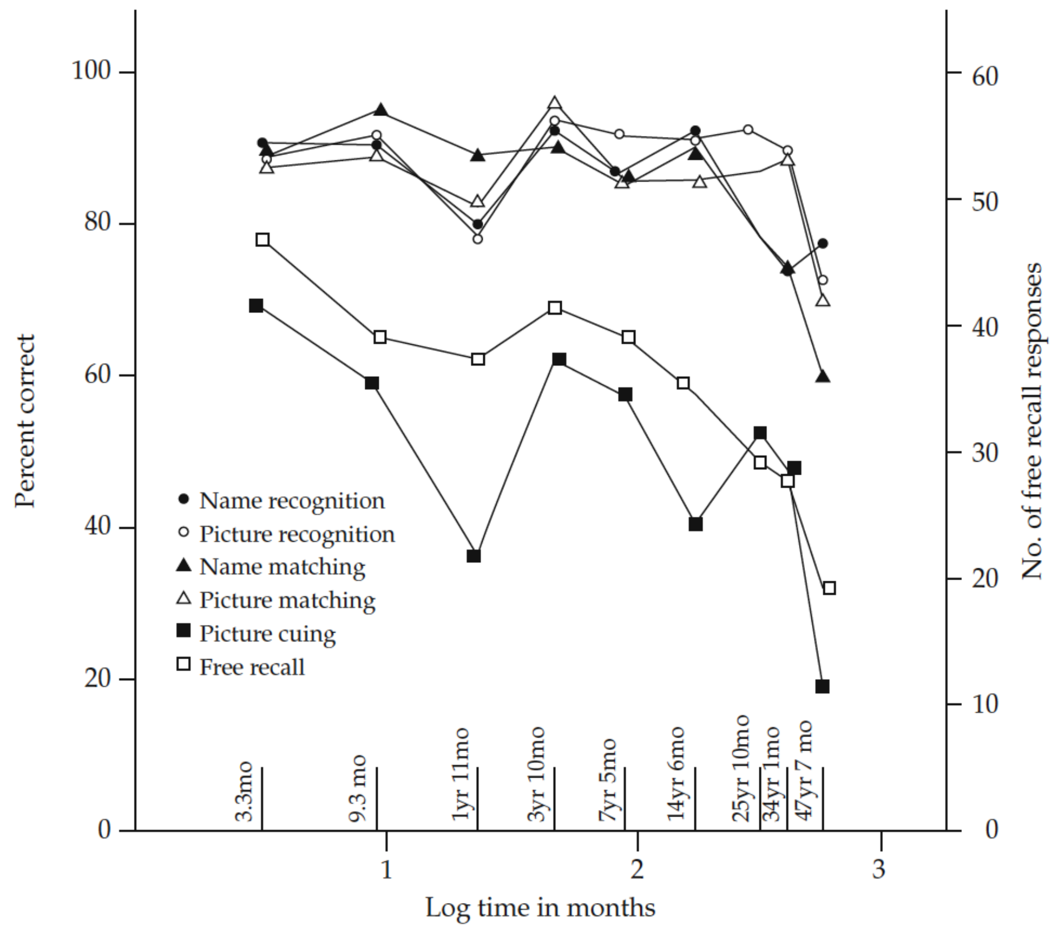

Figure A.20. Retention of names and faces for high school students over time (Bahrick et al. 1975).

The most well-known of the Bahrick studies are those of Spanish language training retention over a 50-year period. A cross-sectional research design using 733 subjects who had received varying levels of language training in high school and college (e.g., single course to multiple years of advanced courses) participated. Supplementary information was also collected, such as grades received and various measures of language use over time (these were low and showed no effects). A prototypical result of this study is shown in Figure A.21.

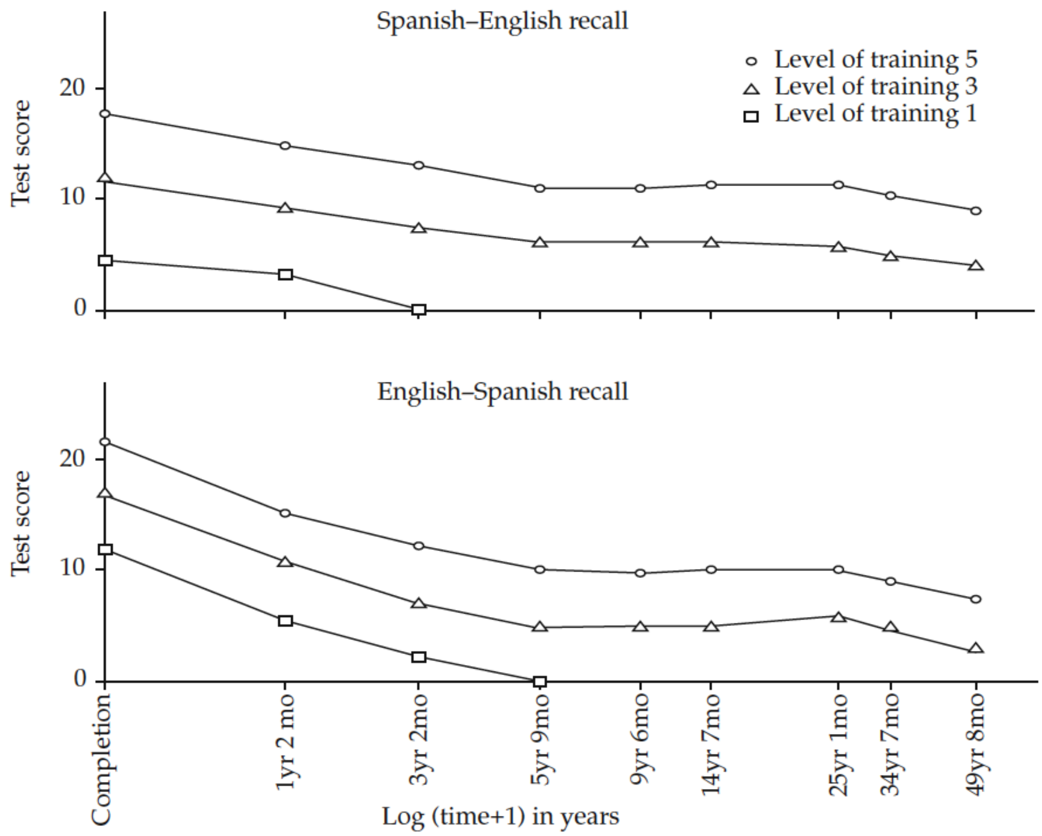

Figure A.21. Vocabulary recall over time as a function of original training level (Bahrick 1984).

The figures show a decline in retention over the first 5 years since original learning, followed by a long period of memory stability up to 35 years, with a further decline after 50 years that is interpreted as age-related. The differences in retention for Spanish language and high school student names and faces during the first five years of retention are related to the recency of content. In language classes, new material is added up to the end of training, with fewer corresponding opportunities to practice. In high school classes, few new students are added over the 4-year period of attendance, so there would be many opportunities to overlearn on a spaced practice basis. An additional important finding from this work is the effect of course grade upon memory retention—the higher the grade, the better the retention. Further, the more training that was initially received, the higher the overall level of retention.

A number of other studies of a similar cross-sectional nature have shown the following effects:

- Retention of mathematical content is relatively high and stable across time for students who have gone beyond the calculus level (Bahrick and Hall 1991a).
- Specific street names and spatial layout of streets in a well-learned city are forgotten over 3–6 years, whereas recognition memory for landmarks stays intact for a much longer period (Bahrick 1983).
- Academic college course content in cognitive psychology shows a several-year period of decline followed by a long period of stability (Conway et al. 1991).

- Basic science knowledge in medical education declines during the course of physician education, but at least 25% of that knowledge remains after 50 years (Custers and ten Cate 2011).

Conway et al. (1992) reviewed the Bahrick and related studies, and concluded that naturalistic knowledge such as memory for personally acquired information—such as knowledge of places or friends—is less prone to sharp initial declines. Instead, there is a gradual diminution of this knowledge over time. Academic knowledge, on the other hand, tends to show somewhat sharp declines for several years prior to leveling off for a greater than 35-year period.

A.5.7.4 Enhancing Access to Marginal Knowledge and Optimizing Practice Schedules

As the ability to recall information declines initially and then stabilizes at sub-mastery levels, it is of interest to know the extent to which previously learned material can be "boosted" to higher levels of recall. Many applied circumstances involve learning a body of material that is not for a sometimes-lengthy period. In a series of investigations, the Bahrick group evaluated what they termed "marginal knowledge," i.e., previously learned content that could no longer be recalled but could be recognized. Their interest was the extent to which additional exposure to that material, what they termed "corrective and preventive maintenance," could re-establish the ability to recall the seemingly forgotten material. This work is somewhat different from the studies of the testing effect discussed previously since that work focused on the impact of successful memory retrievals upon subsequent retention. In studies of marginal knowledge access, the focus is on information that was learned earlier but currently not recallable and which is associated with varying "feelings of knowing."

In two separate studies using vocabulary, picture recognition of famous individuals, and general knowledge questions, Bahrick and Phelps (1987) showed that a single, brief presentation (either by recognition on a multiple-choice test or relearning) of an item that was not recallable on an initial test greatly enhanced the ability to recall that same material several days later—this was referred to as "corrective maintenance." Bahrick and Hall (1991b) showed that a simple recall test (without feedback) of previously learned general material enhanced retention at intervals of two hours and 1 month. What is particularly interesting about this finding is that items that were not successfully recalled on the initial test—referred to as "marginal knowledge"—were better recalled on later tests (referred to as an "upward fluctuation"), and items that were recalled on the initial test were less likely to be forgotten on later tests ("downward fluctuation"). By calculating the impact of these manipulations on individual item recall probability, they were able to devise a "preventive maintenance index" expressing the degree of likelihood that a single test would stabilize memory or enhance it on later tests; it was also shown that while some degree of memory decline was evident over a one-month period, the overall effect was still large.

Subsequent work by this group investigated the combined effects of multiple training sessions and the spacing of those training sessions upon retention over a 5-year period. Bahrick et al. (1993) studied the rate of learning and subsequent retention of foreign language vocabulary pairs using either 13 or 26 learning sessions, which were separated in time by 14, 2,8 or 56 days. Vocabulary items were learned to a criterion of one correct recall of the item, after which the next learning session occurred according to the inter-session interval to which the subject was assigned. At the extreme, the 26-session × 56-day interval learning condition took over 4 years to complete. Testing for retention occurred at the end of the specified training interval. The results of this study are shown in Table A.2, illustrating a clear advantage of both number of training sessions and wider spacing of the training intervals.

Table A.2. Mean percentage recall as a function of number of sessions and interval between sessions over a 5-year period (Bahrick et al. 1993).

Interval between Learning Sessions	13 Learning Sessions	26 Learning Sessions
14 days	43	56
28 days	50	68
56 days	57	76

The general conclusion from this research is that it may be feasible to design refresher training schedules based on knowledge of memory loss rate of previously learned material. An early approach to this was to use the regression equations developed from the cross-sectional research approach (Bahrick 1979). This allowed predictions of, for example, the number of visits to a town that would be necessary to maintain the ability to recall street names.

More recently, researchers have addressed the effects of successive relearning sessions in relation to the original level of learning mastery (e.g., Rawson and Dunlosky 2011) and the application of adaptive practice schedules for language learning (Lindsey et al. 2014). Rawson and Dunlosky (2011) extrapolated their laboratory findings to a typical semester-long course. They recommended a 3+3 approach: practice initially to 3 successful recalls of material at initial learning, followed by 3 relearning sessions spaced throughout the course. This would necessarily involve multiple relearning (study) sessions for different material at different times. Lindsey et al. (2014) tested an approach somewhat like this comparing massed practice with two variants of spaced practice. generic and personalized, for language learning in 8th grade students. Massed practice occurred at the end of a single chapter of material. Generic space practice was selected based on the previous chapter studied, emphasizing earlier material from the chapter. The personalized spaced practice used a quantitative model that inferred the instantaneous memory strength of each item studied; this provided a much more distributed selection of material and practice across chapters. Final test results showed the advantage of the personalized approach to spacing practice, as seen in Figure A.22. The advantage of personalized spacing diminished over time as the interval between initial study and final test decreased, which is a common finding in studies of the spacing effect. There was little difference between massed end-of-chapter practice versus spaced practice of items from the previous chapter, suggesting that insufficient time had elapsed between initial study and practice in this condition for any beneficial effect of the spacing effect.

The personalized spacing model was based on a complex model involving item difficulty, student ability, study history and functions describing the effects of forgetting and practice over time.

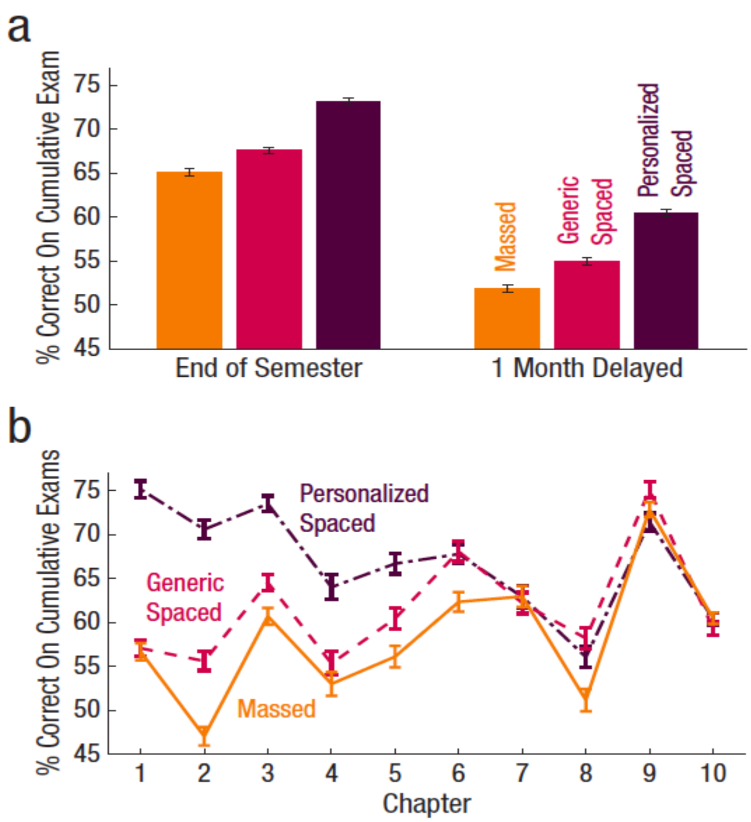

Figure A.22. (a) Percent correct performance on end-of-semester and 1-month post-semester exams for massed, generic spaced and personalized, spaced practice. (b) Percent correct on pooled end-of-semester and 1-month delayed exams for material in each chapter studied (Lindsey et al. 2014).

A.5.7.5 Educational Implications

The results described in this section are generally encouraging from the overall standpoint of the educational system—what is taught is retained at substantial levels for as long as 35 years without rehearsal or relearning. Relearning using long spacing intervals between sessions enhances the ability to recall information many years later. Further, improvements in the concepts of feedback and adaptive study suggest that it may eventually be possible to develop specific algorithms that can predict appropriate schedules of relearning for knowledge maintenance.

A.5.7.6 Relevance for NDE Training

This section provides clear evidence that learned material can be maintained for extensive periods of time up to 35 years, with little or no use or rehearsal. This knowledge is far from perfect—the ability to recall specific material is much lower than the ability to recognize that same material. The best estimates come from the studies of language retention, wherein recall declines over 5 years and stabilizes around 40% of original learning; recognition stabilizes in the 70–80% range.

This general pattern of results has implications for continuing training in NDE. First, the studies suggest that longer periods of training, i.e., continuing to advanced levels, lead to higher overall levels of retention. Thus, a general training strategy for continuing education should emphasize progression from more basic to more advanced topics and/or procedure complexity. In practice, this means that individuals attaining Level II certification should not only maintain that certification but should study toward Level III—this will help to consolidate their more basic knowledge.

Next, the data show that better performance levels during learning, i.e., course grades are associated with better overall very long-term retention. While individuals do differ in capability, this factor should be borne in mind by trainers and managers—all Level IIs or Level IIIs are not the same. PDI testing at EPRI tends to obscure these performance differences, which may yield important information for remedial purposes.

Finally, the research addressing corrective and preventive maintenance of access to knowledge suggests potential improvements in the approach to continuing education and practice. The fact that a simple test can lead to enhanced recall later of material that was not successfully recalled on the original test indicates an important role for continuous testing as a preventive maintenance approach. This might be used with procedure elements, for example, or for specific weld flaw recognition—there are many possibilities given the ability to simulate, store and retrieve flaw patterns. Additionally, for areas where memory is sub-optimal, the corrective maintenance approach of brief relearning sessions has been demonstrated.

A.5.8 Transfer of Training

A.5.8.1 Background

Transfer of training occurs when original learning of material or a task influences subsequent learning or performance of a different task. In this general sense, transfer occurs all the time—our prior school learning, technical training, interactions and general experience of the world influence how we respond to newly encountered situations. Transfer can be either positive (new tasks or materials are learned more quickly or performance is better), negative (prior learning seems to impede subsequent performance on a different task), or neutral (McGeoch and Irion 1952).

An example of positive transfer would be the general skill of driving that allows vehicle control using right-hand drive when original learning took place with left-hand drive. Negative transfer, using the same example, occurs when drivers in right-hand drive vehicles use the stalk on the left of the steering column to signal a turn—this usually activates the windshield wipers in contemporary vehicles as the turn signal is mounted on the right side. This is a common error for the first several days of driving; there is negative transfer in this case because in left-hand drive vehicles (the original training situation) the turn signal is mounted on the left of the steering column.

Knowledge about the general principles of transfer is important because technical training assumes that knowledge will transfer positively to circumstances encountered during operational situations. Further, specific methods of training, such as use of simulators, are based on the assumption that skills developed in lower-fidelity models will apply directly and positively to the work setting. Issues in this regard concern the optimal level of fidelity, whether to train parts of a complex task, the order of training the parts, or whether training the whole task is preferred.

Specific demonstration of transfer is not straightforward in many cases, and there is debate concerning the extent to which knowledge applies across various problem-solving domains

(Barnett and Ceci 2002). There are fundamental concepts, however, that are important to consider in the context of designing and evaluating training programs. This section will discuss some of these concepts and theories and illustrate some recent results concerning transfer of knowledge through test-enhanced learning. This section provides a basis for more specific-analysis training programs in applied settings that are discussed in the next chapter.

A.5.8.2 Theoretical Aspects of Transfer

The earliest theory of transfer is found in the doctrine of formal discipline. This theory of learning asserts that study of one type of academic knowledge, e.g., Latin, will develop mental skills that are applicable in other, very different areas such as mathematics. The basis for this thinking was a "faculty psychology" in which the mind is composed of specific abilities, such as reasoning, which if exercised will improve performance on all tasks dependent on that faculty (Woodworth 1938).

A more technical definition of transfer is based on the concept of stimulus generalization. This is a phenomenon originally observed in Pavlovian classical conditioning, in which the conditioned response could be evoked by a similar, but not identical, stimulus, such as a tone or a bell, to that which was originally used in learning. Stimulus generalization is a robust finding within the context of very specific trained responses and follows a gradient, as shown in Figure A.23.

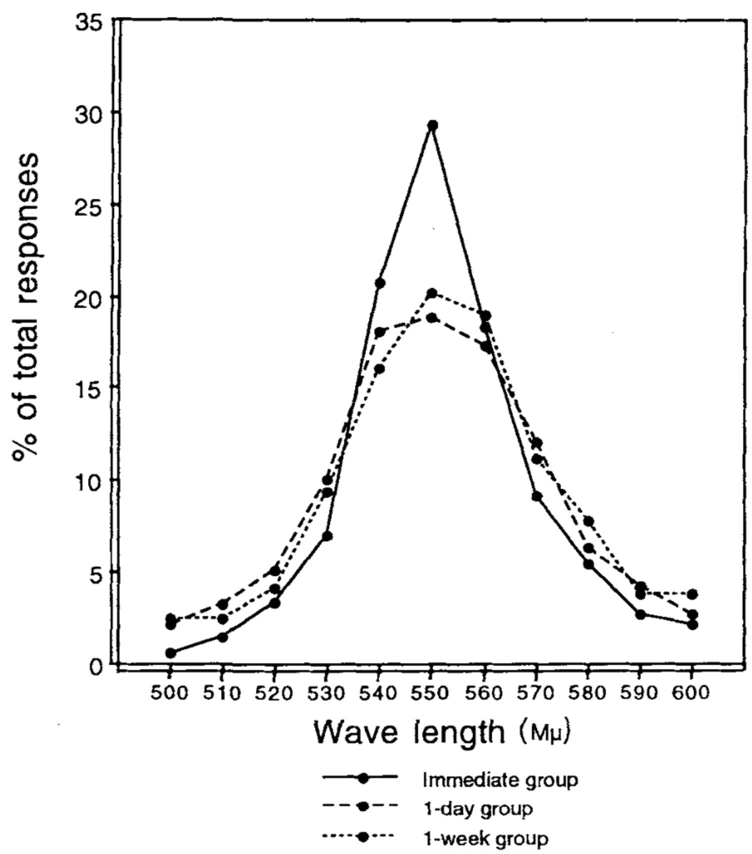

Figure A.23. Percentage of responses by pigeons conditioned to a 550-nanometer stimulus at various delays (Thomas and Lopez 1962).

The results of generalization studies suggest that transfer of learned responses will be better as the degree of similarity with learning conditions increases. Stimulus generalization has been observed across a wide range of learning conditions in both animals and humans (Bouton et al. 1999).

Early experimental tests of transfer were performed by Woodworth and Thorndike (1901). These studies examined, for example, estimating the area of various shapes. The findings indicated that estimation skill for one type of shape did not transfer to another. Such findings were interpreted to mean that “the spread of practice occurs only where identical elements are concerned in the influencing and influenced function.” The Theory of Identical Elements remains the most influential position about transfer and has been re-purposed in various ways to accommodate new experimental findings.

More recent formulations of the identical elements theory include Transfer-Appropriate Processing (Morris et al. 1977) and Procedural Reinstatement (Healy et al. 1992). Transfer-appropriate processing involves using encoding (learning) strategies that will be employed during testing. Morris et al. (1977) showed that paying attention to semantic content generally yields superior recognition memory and recall performance. However, if the memory test requires phonemic judgments such as rhyme then attention to the rhyming aspects of material during initial learning is better. In related research, Healy et al. (1992) found that procedural learning on tasks such as letter detection and spatial memory was well retained over time. For example, students showed much better performance in remembering the locations of buildings in which they took classes, than information concerning time of day or instructor. The effect is explained as a result of original learning using various mental and physical procedures repeatedly (e.g., walking to class) that are then re-invoked at the time of test.

A.5.8.3 Facilitating Transfer

Since transfer occurs on tests of memory for previously learned material, the various factors discussed in previous sections are important for facilitating transfer. These encompass total time spent learning including repetitions; as well as variation in practice, testing, and feedback. Combining these factors in appropriate ways will tend to facilitate transfer of original learning.

An example is provided by recent research by Butler et al. (2017). In this study, the facilitating effects of testing concept learning was evaluated by first exposing learners to conceptual material by reading short paragraphs. Concept application questions were then given. In one condition, each concept was tested with the same question three separate times, while in another condition the concept was tested with three different applications (ordering was mixed in both conditions so that multiple concepts were tested). Subsequently, in a second session learners were asked different questions requiring use of the concepts. The findings showed that testing original learning with variable questions requiring use of the same concept resulted in superior transfer performance on the final test. Butler et al. (2017) suggest that the variability of concept usage introduced during original learning led to generalization in the use of concepts at the final test. These authors review a variety of other studies showing generalization effects in complex human learning based on variation in the original material presented.

A fundamental distinction in research concerning transfer is between *near* and *far* transfer. Near transfer refers to situations that are highly similar to the learning environment, whereas far transfer addresses circumstances that may call upon fundamental knowledge and skills in a context that is highly different from the original learning situation or at a very long time after original learning (Barnett and Ceci 2002). Most of the research we have discussed in this chapter has addressed *near transfer*, i.e., tests of memory for items that were learned in laboratory settings with relatively short retention intervals.

A comprehensive theory of transfer, and hence how to train, must address the entire range of transfer situations. Barnett and Ceci (2002) distinguish between knowledge content (what is learned) and knowledge context (the circumstances of learning and application). Their research framework suggests potential evaluation dimensions for training programs, including the skill that is learned (procedure, representation, principle or heuristic), the performance changes desired (speed, accuracy, strategy), and memory demands (execute an action, recognize a situation and execute, recall-recognize and execute). They further outline a range of contexts that can be used to compare learning situations with application requirements.

A.5.8.4 Educational Implications

A concrete statement of what is best for transfer between educational settings and the world-at-large is not possible based on the current state of research knowledge. Unlike the more tractable issues of how best to facilitate memory for specific material, transfer is much more complex. Wide ranging views of the transfer problem encompass the entire structure of the educational system, from earliest potential “tracking” of students into technical areas or professional preparation to methods for teaching fundamental and more advanced material. At one extreme are the adherents of “situated learning,” which maintains that learning should occur in the context for which it will be applied; at the other are contemporary adherents of the formal discipline, which focuses on training reasoning or critical thinking skills. The use of analytic frameworks such as that of Barnett and Ceci (2002) can be helpful in evaluating specific instances of learning and application.

A.5.8.5 Relevance for NDE Training

Knowledge transfer is an issue that was reflected in the task analysis study of manual ultrasonic inspection. The concern was expressed that entry-level personnel were not provided with adequate field experience opportunities—specifically, that performing flow accelerated corrosion inspections was not appropriate experience for understanding and applying manual UT for weld inspections. More generally, the progression from basic principles in classroom training to use of that knowledge in evaluating potential flaws is a classic problem in knowledge transfer. Are the principles learned in training applied? If so, how? Are there better ways to illustrate—and hence transfer—knowledge of material degradation through operations to observed effects in technical inspections?

The theoretical framework provided by Barnett and Ceci (2002) and the detailed analysis of variables potentially affecting transfer provided by Reder and Klatzky (1994) can be applied to NDE training. Developing specific dimensions for assessing the basic transfer models underlying NDE training may help to identify areas where the use of testing based on examples may be appropriate. Many other possibilities exist for applying the basic research knowledge and analytic concepts of transfer to NDE training. This area will be further discussed in the subsequent chapter addressing technical training and instructional psychology.

A.5.9 Metacognition and Self-Regulated Learning

Metacognition refers to the general set of processes employed by people to monitor their own learning and memory processes—essentially, “thinking about thinking.” These processes are applied during acquisition of information, retention, and retrieval (Dunlosky et al. 2007). There is an interaction between the processes employed to learn, retain and retrieve (control processes) and those used to monitor them. Based on various self-judgments, individuals adapt their control strategies, e.g., by studying more or less, focusing on particular source material, and practicing.

This section briefly describes some key interactions between metacognitive judgments and performance. Self-judgments play a large role in how individuals allocate their study time and strategies, particularly when refresher training and practice is self-guided, as with NDE.

Research concerning metacognitive influences upon self-regulated learning has been reviewed by Bjork et al. (2013). The have found that students do tend to advocate certain effective study techniques, such as self-testing, allocating study time and spaced practice. However, these belief systems are not necessarily put into practice. Bjork et al. (2013) suggest that this is because of metacognitive “illusions” that can lead to erroneous judgments of learning. For example, learners tend to be overconfident in their judgments of the future ability to recall material if they are able to recall it on a current test. On the other hand, they have lower confidence that future learning trials would lead to better performance on material not recalled currently. This is termed “stability bias” and tends to inhibit studying since people think that their current level of knowledge will not degrade or improve with further learning opportunity.

A key factor in making judgments of learning—and hence study decisions—is immediate “fluency” with the material of interest. Fluency is measured in a variety of ways, such as ease of recall, speed of recognition, recency of having seen, etc. Fluency tends to reinforce the illusion of knowledge stability and may thus reduce the tendency to study. It can also reinforce the tendency to mass practice time, when in reality spaced practice leads to better long-term retention. Thus, the ease with which people are currently able to recall material is not necessarily the best indicator of long-term retention. Other factors come into play, such as the amount of distributed practice and the extent to which the material has been tested previously.

In addition to metacognitive biases, Bjork et al. (2013) describe a number of attitudinal and social variables that can adversely affect decision making about learning. In a previous section we described the learning opportunities provided by errors, particularly if associated with informative feedback. More errors will be made under certain study conditions, such as spaced practice, which in turn can lead to better retention. However, errors are typically thought of as something to avoid during learning. Similarly, attribution of differences in performance to innate abilities and the general assumption that learning can and should be easy reduces the perceived importance of training, practice and experience. Bjork and Bjork (2011) described the concept of “desirable difficulties,” which can involve more effort (and mistakes) in the short run but better overall performance in the long run.

A.5.9.1 Educational Implications

The paradox of metacognition is that what learners think they know may not be, and this affects their current and future study habits. The traditional structure of classroom learning and technical training tends to reinforce inaccurate beliefs about how well students may know something by providing massed practice and immediate testing. While it is unlikely that any major change will ensue in these practices in the near future, knowing about the counterproductive beliefs and metacognitive judgments can help both teachers and students to address them. At the level of structuring and presenting material, various practices discussed in previous sections would be appropriate, such as spacing and interleaving subject matter, with testing provided at intervals to boost retention. Performance on these tests and the overall feelings of knowing tend to guide study efforts because these can lead to inaccurate assessments; thus, developing counter-guidelines would be appropriate. For example, students could be encouraged to study all tested material regardless of performance but allocate study time based on need.

A.5.9.2 Relevance for NDE Training

Belief systems and metacognitive judgments are highly relevant to NDE, particularly annual training. Because of the very broad guidance provided by the requirement ("Personnel shall practice UT techniques by examining or by analyzing prerecorded data from material or welds containing flaws similar to those that may be encountered during inservice examinations") there is ample room for individuals to decide they don't need to practice on particular kinds of welds because they feel as if they know them. This may be based on recent experience, a belief that they would not encounter such flaws, or many other factors. Development of guidance to allocate practice time (as well as testing upon that practice) using the basic and theoretical work described above would be appropriate.

A.6 Summary and Principles for Application of Research and Theory

This appendix has reviewed research from the science of human learning and memory as it pertains to the core issues in NDE training—amount of time spent in the field to develop skill, the progressive loss of skill over periods of disuse, and the importance of specific learning factors such as practice, testing, and feedback. Additionally, issues related to very long-term retention, transfer of training, and metacognition provide a basis for interpreting current practices in NDE training. In this summary section, we review the key points of prior sections in the context of what it might tell us about implementing NDE training and articulate core principles of human learning and memory developed from laboratory studies.

The earliest studies of human learning and memory and research up to the present day confirm that learning takes time, that forgetting is a simultaneous process, and that balancing these two processes can help to enhance retention of learned material. Thus, the total time hypothesis of learning is still the best principle to use when considering the structure of training curricula, i.e., **more time spent learning will result in better and longer retention of information**. This principle applies in very short retention intervals as well as over multiple decades. Numerous intervening factors can facilitate how learning time is spent, which can also be incorporated in training implementation.

The body of research reviewed indicates the role of multiple types of memory in NDE, including short-term working memory, long-term declarative memory for factual information, long-term episodic memory for events, procedural memory for sequences, spatial memory for plant layouts, and perceptual-representational memory for specific patterns observed in the past. With the development of expertise through experience, working memory can grow in capacity so that more senior personnel can hold more relevant information about an inspection in their immediate awareness. The inter-relationships between these types of memory and training for NDE in operations is complex; the key principle resulting from this research is that **multiple modes of learning and memory contribute to acquiring, maintaining and applying NDE skill, and that operations experience and classroom laboratory learning are complementary rather than interchangeable.**

This distinction between learning and performance shown by latent and incidental learning indicates that measurement of knowledge and skill by performance tests do not completely represent what is known. Simple exposure and exploration within an environment (latent learning) or manipulating information for some other purpose (incidental learning) can result in task-relevant knowledge that can be applied when conditions warrant. Specific relevance of this general finding concerns how experience hours are gained by NDE trainees—learning may be taking place simply by helping a more experienced examiner. The general principle derived from this research is that **learning occurs in the absence of specific performance demonstrations or requirements, and that performance testing gives an incomplete portrait of the extent of knowledge.**

Overlearning—study of material beyond the point of mastery in close succession to the original study period—is a common characteristic of practical education and training. Research has shown that when the test period is relatively close in time to the learning events, performance is enhanced in contrast to learning to simple mastery, i.e., one perfect test. However, if testing is conducted at later points in time, the effects of overlearning are much diminished. There is some evidence that motor skills, such as equipment assembly/disassembly, results in longer-term effects from overlearning. Conceptual and verbal material, however, tend to degrade over time despite overlearning. Research indicates that overlearning across separate sessions leads to enhanced performance; by spreading out the study time that leads to mastery, retention is enhanced in later performance tests. A more analytic view of overlearning suggests that the underlying principle is that **short-term overlearning results in better immediate test performance but is not durable. Overlearning associated with repeated exposure to foundational material to learn more complex concepts and operations entails spacing of repetition and leads to more durable learning for the long term.**

Studies of repeated learning and practice effects in humans have focused on the distribution and variability of the practice sessions. The previously summarized research on overlearning illustrates that for longer periods, spacing of the learning sessions leads to better retention. Many years of parametric study of the spacing effect in human learning suggests that there is an interaction between spacing intervals and retention intervals; the beneficial impact of spacing depends on the length of time between learning and test. Further, varying how study is performed by using differing versions of the material or types of problems—known as *interleaving*—can improve performance beyond simply repeating the same material that has been presented before. The general principle resulting from this work is that **practice distribution and variability leads to more durable learning than concentrated practice on identical material; longer intervals between practice sessions are appropriate for longer retention intervals, such as several months between practice sessions for testing at a one-year retention interval.**

The research concerning repetition and variability of learning sessions shows that the effective use of spacing and variability in exposure to the material to be learned is beneficial. Further analysis comparing repeated study with the use of testing for previously learned material indicates that testing can enhance memory even more than re-study. There are two ways that this occurs: (1) by re-focusing learner attention on material for which they perform poorly, and (2) a direct effect on the memory trace. The latter effect is of primary interest in this review since there are other means of re-focusing attention. The “testing effect” has been known for many years in educational settings, and laboratory studies have revealed that while repeated study is more beneficial for immediate test performance, testing—with or without feedback—produces better retention over longer periods. Related findings indicate that when learners engage in self-explanation of relationships in studied material or questions are asked regarding relationships, recall is better. In general, **testing of previously learned material is preferred over simple re-study, and testing that involves elaborating semantic linkages within the material can potentiate retention.**

The concept of feedback in learning can be traced to reinforcement, a consequence of behavior that increases its likelihood. Feedback can come directly from muscle action (intrinsic) or from the results of task performance (extrinsic). In cognitive tasks, feedback is generally extrinsic and can be delivered either immediately or at some delay. Further, feedback can provide varying levels of information, such as correct-incorrect, hints regarding the correct answer (known as scaffolded feedback), and correct answer. Voluminous studies in a wide range of tasks show gradations in performance depending on how and when feedback is provided. It is generally the case in both laboratory and traditional classroom settings that **retention of information is best when delayed correct answer feedback is provided.** The delay in laboratory setting is typically provided by structuring the learning and test sessions so that feedback is provided after all items have been tested, while delay of a day or more shows beneficial effects in classroom settings. Further, low-confidence responses can be reinforced, while high confidence errors can be reduced via correct answer feedback. Studies of error generation during learning suggest that the general philosophy that errors should be minimized during learning is both unrealistic and incorrect—instead, **errors represent learning opportunities when paired with appropriate feedback and can specifically enhance learning.**

Most of the research about human learning and memory, both in laboratory and in classroom settings, address retention over a relatively short period of time of at most 1 year. The few studies of very long-term retention, i.e., multiple decades, have used cross-sectional research methods to assess knowledge maintenance over long periods of time. The general finding from these studies is that **knowledge retention declines over the first five years after learning, then stabilizes at levels around 40% for as long as 35 years**. Correlation of very long-term retention with original learning factors such as grades received and level of training suggest that **better performance in original learning and greater levels of advanced training lead to better long-term knowledge maintenance.** Further laboratory studies using shorter retention intervals suggest that periodic relearning for “marginal knowledge” serves a preventive maintenance function and increases likelihood of retention for longer periods. Specific **studies of relearning schedule optimization suggest that providing additional exposure to material at intervals where knowledge has decayed somewhat, but prior to final testing, is beneficial for longer-term performance.**

Transfer of training occurs when what is originally learned is applied in the performance of a different task. This is essentially a generalization of prior learning to a situation with some degree of similarity to the original. Transfer can be positive, i.e., the new task is performed better than if no original learning had taken place, or negative, i.e., the new task is performed more poorly. Substantial research has been conducted to determine the relationship between original learning and subsequent application. The general finding is that **to the extent that there is similarity between the new task and original learning, positive transfer occurs. As the degree of similarity differs, or if the new task involves opposite response patterns from what was learned, negative transfer occurs.** Studies of training transfer are particularly relevant to issues of appropriate experience for professional qualification and the use of simulators in developing skills.

A key set of factors that influence learning and retention are individual characteristics, such as self-assessments of what you know and how well you know it. The general findings from studies in this area are that **self-assessments of knowledge are inaccurate and tend to reinforce**

ineffective learning strategies such as re-reading. People will tend to avoid using learning strategies such as spaced practice and self-testing in favor of intuitively appealing and relatively easier strategies such as review and re-reading. Research indicates that **"desirable difficulties" in learning that may entail more mistakes in the short run will pay dividends of better retention in the long run.**

With respect to NDE training, the principles summarized above suggest several approaches, albeit somewhat idealized, to implementation. These would include:

- A programmed learning approach to classroom/text-based material that uses appropriately spaced and elaborated repetitions, as well as testing with feedback provided at the end of the test. Standardization of this material is recommended to reduce variation across training providers.
- Interleaving of classroom/text-based material so that subject-matter variety is provided.
- A period of field experience commensurate with observing and assisting with the types of exams to be performed upon certification.
- Practice in summarizing findings from field experience assistance to develop expert working memory capability.
- Performance demonstration tests that focus on more than a single procedure or weld type to promote variability in practice and retrieval-based testing.
- Development of a means to provide some type of feedback for performance testing that is more informative than pass/ no-pass.
- Development of a catalogue of common errors and feedback that is focused on correcting inspection process elements that engender those errors.
- Requirement for refresher training that is based on the expected intervals between knowledge use and expected decay.
- Requirement for testing of refresher training, possibly developed in a manner so that periodic testing/training can be performed concurrently.

Additional review material described in Appendix B focusing on technical training, instructional psychology, and domain expertise will be used to elaborate and contextualize the fundamental principles developed in this volume.

A.7 References

10 CFR 50. 2018. "Incorporation by Reference of American Society of Mechanical Engineers Codes and Code Cases." *Code of Federal Regulations*, U.S. Nuclear Regulatory Commission, Washington, D.C. Available at https://www.nrc.gov/docs/ML1613/ML16130A530.pdf.

Adams JA. 1968. "Response Feedback and Learning." *Psychological Bulletin* 79(6):486-504. DOI: 10.1037/h0026741.

Adams JA. 1987. "Historical Review and Appraisal of Research on the Learning, Retention, and Transfer of Human Motor Skills." *Psychological Bulletin* 101(1):41-74. DOI: 10.1037/0033-2909.101.1.41.

Atkinson RC and RM Shiffrin. 1968. "Human Memory: A Proposed System and Its Control Processes." In *The Psychology of Learning and Motivation: Advances in Research and Theory*, pp. 89-195, ed: KW Spence. NY: Academic Press.

Baddeley A. 1992. "Working Memory." *Science* 255(5044):556-559. DOI: 10.1126/science.1736359.

Baddeley A. 2012. "Working Memory: Theories, Models, and Controversies." *Annual Review of Psychology* 63(1):1-29. DOI: 10.1146/annurev-psych-120710-100422.

Baddeley AD. 1999. *Essentials of Human Memory*, East Sussex, United Kingdom: Psychology Press, Taylor and Francis Group.

Baddeley AD and DJA Longman. 1978. "The Influence of Length and Frequency of Training Session on the Rate of Learning to Type." *Ergonomics* 21(8):627-635. DOI: 10.1080/00140137808931764.

Bahrick HP. 1979. "Maintenance of Knowledge: Questions about Memory We Forgot to Ask." *Journal of Experimental Psychology: General* 108(3):296-308. DOI: 10.1037/0096-3445.108.3.296.

Bahrick HP. 1983. "The Cognitive Map of a City: Fifty Years of Learning and Memory." In *The Psychology of Learning and Motivation: Advances in Research and Theory*, pp. 125-163, ed: G Bower. New York: Academic Press.

Bahrick HP. 1984. "Semantic Memory Content in Permastore: Fifty Years of Memory for Spanish Learned in School." *Journal of Experimental Psychology: General* 113(1):1-29. DOI: 10.1037/0096-3445.113.1.1.

Bahrick HP. 2000. "Long-Term Maintenance of Knowledge." In *The Oxford Handbook of Memory*, pp. 347-362, eds: E Tulving and FIM Craik. New York, NY: Oxford University Press.

Bahrick HP, LE Bahrick, AS Bahrick and PE Bahrick. 1993. "Maintenance of Foreign Language Vocabulary and the Spacing Effect." *Psychological Science* 4(5):316-321. DOI: 10.1111/j.1467-9280.1993.tb00571.x.

Bahrick HP, PO Bahrick and RP Wittlinger. 1975. "Fifty Years of Memory for Names and Faces: A Cross-Sectional Approach." *Journal of Experimental Psychology: General* 104(1):54-75. DOI: 10.1037/0096-3445.104.1.54.

Bahrick HP and LK Hall. 1991a. "Lifetime Maintenance of High School Mathematics Content." *Journal of Experimental Psychology: General* 120(1):20-33. DOI: 10.1037/0096-3445.120.1.20.

Bahrick HP and LK Hall. 1991b. "Preventive and Corrective Maintenance of Access to Knowledge." *Applied Cognitive Psychology* 5(1):1-18. DOI: 10.1002/acp.2350050102.

Bahrick HP and LK Hall. 2005. "The Importance of Retrieval Failures to Long-Term Retention: A Metacognitive Explanation of the Spacing Effect." *Journal of Memory and Language* 52(4):566-577. DOI: 10.1016/j.jml.2005.01.012.

Bahrick HP, LK Hall and MK Baker. 2013. *Life-Span Maintenance of Knowledge*, New York: Taylor and Francis.

Bahrick HP and E Phelps. 1987. "Retention of Spanish Vocabulary Over 8 Years." *Journal of Experimental Psychology Learning Memory and Cognition* 13(2):344-349. DOI: 10.1037//0278-7393.13.2.344.

Barnett SM and SJ Ceci. 2002. "When and Where Do We Apply What We Learn?: A Taxonomy for Far Transfer." *Psychological Bulletin* 128(4):612-637. DOI: 10.1037/0033-2909.128.4.612.

Berry DC and DE Broadbent. 1984. "On the Relationship between Task Performance and Associated Verbalizable Knowledge." *The Quarterly Journal of Experimental Psychology Section A* 36(2):209-231. DOI: 10.1080/14640748408402156.

Bjork EL and RA Bjork. 2011. "Making Things Hard on Yourself, But in a Good Way: Creating Desirable Difficulties to Enhance Learning." In *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, pp. 56-64, eds: MA Gernsbacher, RW Pew, LM Hough and JR Pomerantz. New York, NY: Worth Publishers.

Bjork RA, J Dunlosky and N Kornell. 2013. "Self-Regulated Learning: Beliefs, Techniques, and Illusions." *Annual Review of Psychology* 64(1):417-444. DOI: 10.1146/annurev-psych-113011-143823.

Blodgett HC. 1929. *The Effect of the Introduction of Reward Upon the Maze Performance in Rats*, Berkeley, CA: University of California Publications in Psychology.

Bouton ME, JB Nelson and JM Rosas. 1999. "Stimulus Generalization, Context Change, and Forgetting." *Psychological Bulletin* 125(2):171-186. DOI: 10.1037/0033-2909.125.2.171.

Brown AS. 1991. "A Review of the Tip-of-the-Tongue Experience." *Psychological Bulletin* 109(2):204-223. DOI: 10.1037/0033-2909.109.2.204.

Bryan WL and N Harter. 1899. "Studies on the Telegraphic Language: The Acquisition of a Hierarchy of Habits." *Psychological Review* 6(4):345-375. DOI: 10.1037/h0073117.

Butler AC, AC Black-Maier, ND Raley and EJ Marsh. 2017. "Retrieving and Applying Knowledge to Different Examples Promotes Transfer of Learning." *Journal of Experimental Psychology: Applied* 23(4):433-446. DOI: 10.1037/xap0000142.

Butler AC, JD Karpicke and HL Roediger. 2008. "Correcting a Metacognitive Error: Feedback Increases Retention of Low-Confidence Correct Responses." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 34(4):918-928. DOI: 10.1037/0278-7393.34.4.918.

Butler AC and HL Roediger. 2007. "Testing Improves Long-Term Retention in a Simulated Classroom Setting." *European Journal of Cognitive Psychology* 19(4-5):514-527. DOI: 10.1080/09541440701326097.

Butler AC and HL Roediger. 2008. "Feedback Enhances the Positive Effects and Reduces the Negative Effects of Multiple-Choice Testing." *Memory & Cognition* 36(3):604-616. DOI: 10.3758/mc.36.3.604.

Butterfield B and J Metcalfe. 2001. "Errors Committed with High Confidence are Hypercorrected." *Journal of Experimental Psychology. Learning, Memory, and Cognition* 27(6):1491-1494. DOI: 10.1037/0278-7393.27.6.1491.

Carpenter SK. 2017. "Spacing Effects on Learning and Memory." In *Cognitive Psychology of Memory, Vol. 2 of Learning and Memory: A Comprehensive Reference, 2nd Edition*, pp. 465-485, ed: JT Wixted. Oxford: Academic Press.

Cepeda NJ, H Pashler, E Vul, JT Wixted and D Rohrer. 2006. "Distributed Practice in Verbal Recall Tasks: A Review and Quantitative Synthesis." *Psychological Bulletin* 132(3):354-380. DOI: 10.1037/0033-2909.132.3.354.

Cepeda NJ, E Vul, D Rohrer, JT Wixted and H Pashler. 2008. "Spacing Effects in Learning: A Temporal Ridgeline of Optimal Retention." *Psychological Bulletin* 19(11):1095-1102. DOI: 10.1111/j.1467-9280.2008.02209.x.

Chase WG and HA Simon. 1973. "The Mind’s Eye in Chess." In *Visual Information Processing*, pp. 215-281, ed: WG Chase. NY: Academic Press.

Chrastil ER. 2013. "Neural Evidence Supports a Novel Framework for Spatial Navigation." *Psychonomic Bulletin and Review* 20(2):208-227. DOI: 10.3758/s13423-012-0351-6.

Collins AM and EF Loftus. 1975. "A Spreading-Activation Theory of Semantic Processing." *Psychological Review* 82(6):407-428. DOI: 10.1037/0033-295X.82.6.407.

Conway MA, G Cohen and N Stanhope. 1991. "On the Very Long-Term Retention of Knowledge Acquired Through Formal Education: Twelve Years of Cognitive Psychology." *Journal of Experimental Psychology: General* 120(4):395-409. DOI: 10.1037/0096-3445.120.4.395.

Conway MA, G Cohen and N Stanhope. 1992. "Very Long-Term Memory for Knowledge Acquired at School and University." *Applied Cognitive Psychology* 6(6):467-482. DOI: 10.1002/acp.2350060603.

Cowan N. 2001. "The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity." *Behavioral and Brain Sciences* 24(1):87-185. DOI: 10.1017/S0140525X01003922.

Cowan N. 2015. "George Miller's Magical Number of Immediate Memory in Retrospect: Observations on the Faltering Progression of Science." *Psychological Review* 122(3):536-541. DOI: 10.1037/a0039035.

Craik FIM and RS Lockhart. 1972. "Levels of Processing: A Framework for Memory Research." *Journal of Verbal Learning and Verbal Behavior* 11(6):671-684. DOI: 10.1016/S0022-5371(72)80001-X.

Craik FIM and E Tulving. 1975. "Depth of Processing and the Retention of Words in Episodic Memory." *Journal of Experimental Psychology: General* 104(3):268-294. DOI: 10.1037/0096-3445.104.3.268.

Cumblidge S and A D'Agostino. 2016. "Calculating Inspector Probability of Detection Using Performance Demonstration Program Pass Rates." In *42nd Annual Review of Progress in Quantitative Nondestructive Evaluation: Incorporating the 6th European-American Workshop on Reliability of NDE. AIP Conference Proceedings Vol. 1706*, p. 200013. July 26-31, 2015, Minneapolis, MN. DOI 10.1063/1.4940657. AIP.

Cumblidge SE. 2018. "Training and Certification." Presented at *NRC/Industry NDE Technical Information Exchange Meeting*, January 16-18, 2018, Washington, DC. ADAMS Accession No. ML18012A723.

Custers EJFM and OTJ ten Cate. 2011. "Very Long-Term Retention of Basic Science Knowledge in Doctors After Graduation." *Medical Education* 45(4):422-430. DOI: 10.1111/j.1365-2923.2010.03889.x.

Davis RA and CC Moore. 1935. "Methods of Measuring Retention." *The Journal of General Psychology* 12(1):144-155. DOI: 10.1080/00221309.1935.9920093.

de Groot AD. 1946. *Thought and Choice and Chess*, The Hague, Netherlands: Mouton.

Diamond R and M Moezzi. 2000. "Revealing Myths about People, Energy and Buildings." In *2000 Summer Study on Energy Efficiency in Buildings: Efficiency & Sustainability*, pp. 8.65-8.76. August 20-25, 2000, Pacific Grove, CA. American Council for an Energy-Efficient Economy (ACEEE), Washington, DC. Available at https://aceee.org/files/proceedings/2000/data/index.htm.

Driskell JE, RP Willis and C Copper. 1992. "Effect of Overlearning on Retention." *Journal of Applied Psychology* 77(5):615-622. DOI: 10.1037/0021-9010.77.5.615.

Dunlosky J, KA Rawson, EJ Marsh, MJ Nathan and DT Willingham. 2013. "Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology." *Psychological Science in the Public Interest* 14(1):4-58. DOI: 10.1177/1529100612453266.

Dunlosky J, M Serra and JMC Baker. 2007. "Metamemory Applied." In *Handbook of Applied Cognition, 2nd Edition*, pp. 137-159, eds: FT Durso, RS Nickerson, ST Dumais, S Lewandowsky and TJ Perfect. NY: Wiley.

Eagle M and E Leiter. 1964. "Recall and Recognition in Intentional and Incidental Learning." *Journal of Experimental Psychology* 68(1):58-63. DOI: 10.1037/h0044655.

Ebbinghaus H. 1885/1964. *Memory: A Contribution to Experimental Psychology*, New York: Dover. Originally published in German in 1885.

Erdelyi MH and J Becker. 1974. "Hypermnesia for Pictures: Incremental Memory for Pictures But Not Words in Multiple Recall Trials." *Cognitive Psychology* 6(1):159-171. DOI: 10.1016/0010-0285(74)90008-5.

Erdelyi MH and J Kleinbard. 1978. "Has Ebbinghaus Decayed with Time? The Growth of Recall (Hypermnesia) Over Days." *Journal of Experimental Psychology: Human Learning and Memory* 4(4):275-289. DOI: 10.1037/0278-7393.4.4.275.

Ericsson KA, Ed. 2009. *Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments*. NY: Cambridge University Press.

Ericsson KA and W Kintsch. 1995. "Long-Term Working Memory." *Psychological Review* 102(2):211-245. DOI: 10.1037/0033-295X.102.2.211.

Fazio LK, BJ Huelser, A Johnson and EJ Marsh. 2010. "Receiving Right/Wrong Feedback: Consequences for Learning." *Memory* 18(3):335-350. DOI: 10.1080/09658211003652491.

Finn B and J Metcalfe. 2010. "Scaffolding Feedback to Maximize Long-Term Error Correction." *Memory & Cognition* 38(7):951-961. DOI: 10.3758/mc.38.7.951.

Fitts PM. 1964. "Perceptual-Motor Skill Learning." In *Categories of Human Learning*, pp. 243-285, ed: AW Melton. NY: Academic Press.

Gagne RM. 1962. "Military Training and Principles of Learning." *American Psychologist* 17(2):83-91. DOI: 10.1037/h0048613.

Gagne RM. 1965. *The Conditions of Learning*, NY: Holt, Rinehart and Winston, Inc.

Gates AI. 1917. "Recitation as a Factor in Memorizing." *Archives of Psychology* 6(40):140.

Gerbier E and TC Toppino. 2015. "The Effect of Distributed Practice: Neuroscience, Cognition, and Education." *Trends in Neuroscience and Education* 4(3):49-59. DOI: 10.1016/j.tine.2015.01.001.

Hattie J, M Gan and C Brooks. 2017. "Instruction Based on Feedback." In *Handbook of Research on Learning and Instruction, 2nd Edition*, pp. 290-324, eds: RE Mayer and PA Alexander. NY: Routledge.

Healy AF, DW Fendrick, RJ Crutcher, WT Wittman, AT Gesi, KA Ericsson and LE Bourne Jr. 1992. "The Long-Term Retention of Skills." In *From Learning Processes to Cognitive Processes: Essays in Honor of William K. Estes, Vol. 2*, pp. 87-118, eds: AF Healy, SM Kossly and RM Shiffin. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Healy AF, VI Schneider and LE Bourne Jr. 2012. "Empirically Valid Principles of Training." In *Training Cognition: Optimizing Efficiency, Durability and Generalizability*, pp. 13-39, eds: AF Healy and LE Bourne Jr. NY: Psychology Press, Taylor and Francis Group.

Hout MC and SD Goldinger. 2010. "Learning in Repeated Visual Search." *Attention, Perception & Psychophysics* 72(5):1267-1282. DOI: 10.3758/app.72.5.1267.

Hovland CI. 1951. "Human Learning and Retention." In *Handbook of Experimental Psychology*, ed: SS Stevens. New York: John Wiley & Sons.

Huelser BJ and J Metcalfe. 2012. "Making Related Errors Facilitates Learning, But Learners Do Not Know It." *Memory & Cognition* 40(4):514-527. DOI: 10.3758/s13421-011-0167-z.

James W. 1890. *The Principles of Psychology*, NY: Dover. Originally published by Henry Holt & Co.

Jiang YV, BY Won and KM Swallow. 2014. "First Saccadic Eye Movement Reveals Persistent Attentional Guidance by Implicit Learning." *Journal of Experimental Psychology: Human Perception and Performance* 40(3):1161-1173. DOI: 10.1037/a0035961.

Kang SHK and H Pashler. 2012. "Learning Painting Styles: Spacing is Advantageous When It Promotes Discriminative Contrast." *Applied Cognitive Psychology* 26(1):97-103. DOI: 10.1002/acp.1801.

Karpicke JD. 2017. "Retrieval-Based Learning: A Decade of Progress." In *Cognitive Psychology of Memory, Vol. 2 of Learning and Memory: A Comprehensive Reference, 2nd Edition*, pp. 487-514, eds: JT Wixted and JH Byrne. Oxford: Academic Press.

Kellman PJ. 2013. "Adaptive and Perceptual Learning Technologies in Medical Education and Training." *Military Medicine* 178(suppl_10):98-106. DOI: 10.7205/milmed-d-13-00218.

Kellman PJ and CM Massey. 2013. "Chapter Four - Perceptual Learning, Cognition, and Expertise." In *Psychology of Learning and Motivation*, pp. 117-165, ed: BH Ross. Cambridge, MA: Academic Press. Vol. 58.

Kellogg WN. 1946. "The Learning Curve for Flying an Airplane." *Journal of Applied Psychology* 30(5):435-441. DOI: 10.1037/h0060547.

Kerr R and B Booth. 1978. "Specific and Varied Practice of Motor Skill." *Perceptual and Motor Skills* 46(2):395-401. DOI: 10.1177/003151257804600201.

Knowlton BJ, ALM Siegel and TD Moody. 2017. "Procedural Learning in Humans." In *Memory Systems, Vol. 3 of Learning and Memory: A Comprehensive Reference, 2nd Edition*, pp. 295-312, ed: H Eichenbaum. Oxford: Academic Press.

Kornell N, MJ Hays and RA Bjork. 2009. "Unsuccessful Retrieval Attempts Enhance Subsequent Learning." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 35(4):989-998. DOI: 10.1037/a0015729.

Krueger WCF. 1929. "The Effect of Overlearning on Retention." *Journal of Experimental Psychology* 12:71-78. DOI: 10.1037/h0072036.

Kulhavy RW. 1977. "Feedback in Written Instruction." *Review of Educational Research* 47(2):211-232. DOI: 10.3102/00346543047002211.

Levine B, E Svoboda, JF Hay, G Winocur and M Moscovitch. 2002. "Aging and Autobiographical Memory: Dissociating Episodic from Semantic Retrieval." *Psychology and Aging* 17(4):677-689. DOI: 10.1037/0882-7974.17.4.677.

Lindsey RV, JD Shroyer, H Pashler and MC Mozer. 2014. "Improving Students' Long-Term Knowledge Retention Through Personalized Review." *Psychological Science* 25(3):639-647. DOI: 10.1177/0956797613504302.

Madigan SA. 1969. "Intraserial Repetition and Coding Processes in Free Recall." *Journal of Verbal Learning and Verbal Behavior* 8(6):828-835. DOI: 10.1016/S0022-5371(69)80050-2.

Marsh EJ, PK Agarwal and HL Roediger. 2009. "Memorial Consequences of Answering SAT II Questions." *Journal of Experimental Psychology: Applied* 15(1):1-11. DOI: 10.1037/a0014721.

Mayer RE. 2011. *Applying the Science of Learning, 1st Edition*, San Francisco: Pearson. ISBN 978-0136117575.

Mayer RE. 2012. "Advances in Applying the Science of Learning to Education: An Historical Perspective." *Journal of Applied Research in Memory and Cognition* 1(4):249-250. DOI: 10.1016/j.jarmac.2012.10.001.

McGeoch JA and AL Irion. 1952. *The Psychology of Learning and Memory*, New York: Longmans, Green and Co.

McKeachie WJ. 1974. "The Decline and Fall of the Laws of Learning." *Educational Researcher* 3(3):7-11. DOI: 10.3102/0013189x003003007.

McNamara TP. 2017. "Spatial Memory and Navigation." In *Cognitive Psychology of Memory, Vol. 2 of Learning and Memory: A Comprehensive Reference, 2nd Edition*, pp. 337-355, ed: JT Wixted. Oxford: Academic Press.

Melton AW. 1963. "Implications of Short-Term Memory for a General Theory of Memory." *Journal of Verbal Learning and Verbal Behavior* 2(1):1-21. DOI: 10.1016/S0022-5371(63)80063-8.

Melton AW. 1970. "The Situation with Respect to the Spacing of Repetitions and Memory." *Journal of Verbal Learning and Verbal Behavior* 9(5):596-606. DOI: 10.1016/S0022-5371(70)80107-4.

Metcalfe J. 2017. "Learning from Errors." *Annual Review of Psychology* 68(1):465-489. DOI: 10.1146/annurev-psych-010416-044022.

Metcalfe J and N Kornell. 2007. "Principles of Cognitive Science in Education: The Effects of Generation, Errors, and Feedback." *Psychonomic Bulletin & Review* 14(2):225-229. DOI: 10.3758/bf03194056.

Miller GA, E Galanter and KH Pribram. 1960. *Plans and the Structure of Behavior*, NY: Holt.

Mitchell DB. 2006. "Nonconscious Priming After 17 Years: Invulnerable Implicit Memory?" *Psychological Science* 17(11):925-929. DOI: 10.1111/j.1467-9280.2006.01805.x.

Morris CD, JD Bransford and JJ Franks. 1977. "Levels of Processing Versus Transfer Appropriate Processing." *Journal of Verbal Learning and Verbal Behavior* 16(5):519-533. DOI: 10.1016/S0022-5371(77)80016-9.

National Research Council. 2017. *Building America's Skilled Technical Workforce*, Washington, DC: National Academies Press.

NRC. 2009. *Nondestructive Examination (NDE) Technology and Codes Student Manual. Chapter 2. Personnel Qualification and Training*. Washington, DC: Nuclear Regulatory Commission. ADAMS Accession No. ML12146A173.

Pashler H, NJ Cepeda, JT Wixted and D Rohrer. 2005. "When Does Feedback Facilitate Learning of Words?" *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(1):3-8. DOI: 10.1037/0278-7393.31.1.3.

Payne DG. 1987. "Hypermnesia and Reminiscence in Recall: A Historical and Empirical Review." *Psychological Bulletin* 101(1):5-27. DOI: 10.1037/0033-2909.101.1.5.

Postman L. 1962. "Retention as a Function of Degree of Overlearning." *Science* 135(3504):666-667. DOI: 10.1126/science.135.3504.666.

Pressey SL. 1926. "A Simple Apparatus which Gives Tests and Scores – and Teaches." *School and Society* 23(586):373-376.

Pusic M, M Pecaric and K Boutis. 2011. "How Much Practice Is Enough? Using Learning Curves to Assess the Deliberate Practice of Radiograph Interpretation." *Academic Medicine* 86(6):731-736. DOI: 10.1097/ACM.0b013e3182178c3c.

Pyc MA and J Dunlosky. 2010. "Toward an Understanding of Students' Allocation of Study Time: Why Do They Decide to Mass or Space Their Practice?" *Memory & Cognition* 38(4):431-440. DOI: 10.3758/mc.38.4.431.

Rasmussen J. 1983. "Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models." *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13(3):257-266. DOI: 10.1109/TSMC.1983.6313160.

Rawson KA and J Dunlosky. 2011. "Optimizing Schedules of Retrieval Practice for Durable and Efficient Learning: How Much is Enough?" *Journal of Experimental Psychology: General* 140(3):283-302. DOI: 10.1037/a0023956.

Reder LM and RL Klatzky. 1994. "Transfer: Training for Performance." In *Learning, Remembering, Believing: Enhancing Human Performance*, pp. 25-56, eds: D Druckman and RA Bjork. Washington, DC: The National Academies Press.

Roediger HL and JD Karpicke. 2006. "Test-Enhanced Learning:Taking Memory Tests Improves Long-Term Retention." *Psychological Science* 17(3):249-255. DOI: 10.1111/j.1467-9280.2006.01693.x.

Roediger HL and DG Payne. 1982. "Hypermnesia: The Role of Repeated Testing." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 8(1):66-72. DOI: 10.1037/0278-7393.8.1.66.

Roediger HL and LA Thorpe. 1978. "The Role of Recall Time in Producing Hypermnesia." *Memory & Cognition* 6(3):296-305. DOI: 10.3758/bf03197459.

Roediger III HL and AC Butler. 2011. "The Critical Role of Retrieval Practice in Long-Term Retention." *Trends in Cognitive Sciences* 15(1):20-27. DOI: 10.1016/j.tics.2010.09.003.

Roediger III HL and K Srinivas. 1993. "Specificity of Operations in Perceptual Priming." In *Implicit Memory: New Directions in Cognition, Development, and Neuropsychology*, pp. 17-48, eds: P Graf and MEJ Masson. Hillsdale, NJ: Lawrence Earlbaum Associates.

Rohrer D. 2009. "Avoidance of Overlearning Characterises the Spacing Effect." *European Journal of Cognitive Psychology* 21(7):1001-1012. DOI: 10.1080/09541440802425954.

Rohrer D and K Taylor. 2006. "The Effects of Overlearning and Distributed Practise on the Retention of Mathematics Knowledge." *Applied Cognitive Psychology* 20(9):1209-1224. DOI: 10.1002/acp.1266.

Rohrer D, K Taylor, H Pashler, JT Wixted and NJ Cepeda. 2005. "The Effect of Overlearning on Long-Term Retention." *Applied Cognitive Psychology* 19(3):361-374. DOI: 10.1002/acp.1083.

Rowland CA. 2014. "The Effect of Testing Versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect." *Psychological Bulletin* 140(6):1432-1463. DOI: 10.1037/a0037559.

Rubin DC and AE Wenzel. 1996. "One Hundred Years of Forgetting: A Quantitative Description of Retention." *Psychological Review* 103(4):734-760. DOI: 10.1037/0033-295X.103.4.734.

Salmoni AW, RA Schmidt and CB Walter. 1984. "Knowledge of Results and Motor Learning: A Review and Critical Reappraisal." *Psychological Bulletin* 95(3):355-386. DOI: 10.1037/0033-2909.95.3.355.

Sanquist TF, SL Morrow, AL D'Agostino, NM Hughes and CM Franklin. 2018. *Human Factors in Nondestructive Examination: Manual Ultrasonic Testing Task Analysis and Field Research*. PNNL-27441. Richland, WA: Pacific Northwest National Laboratory. ADAMS Accession No. ML18176A055.

Schachter DL, AD Wagner and RL Buckner. 2000. "Memory Systems of 1999." In *The Oxford Handbook of Memory*, eds: E Tulving and FIM Craik. New York, NY: Oxford University Press.

Shoben EJ. 1992. "Semantic Memory." In *Encyclopedia of Learning and Memory*, pp. 581-585, ed: LR Squire. NY: Macmillan Publishing.

Skinner BF. 1954. "The Science of Learning and the Art of Teaching." *Harvard Educational Review* 24:86-97.

Snodgrass JG and K Feenan. 1990. "Priming Effects in Picture Fragment Completion: Support for the Perceptual Closure Hypothesis." *Journal of Experimental Psychology: General* 119(3):276-296. DOI: 10.1037/0096-3445.119.3.276.

Soderstrom NC and RA Bjork. 2015. "Learning Versus Performance:An Integrative Review." *Perspectives on Psychological Science* 10(2):176-199. DOI: 10.1177/1745691615569000.

Spitzer HF. 1939. "Studies in Retention." *Journal of Educational Psychology* 30(9):641-656. DOI: 10.1037/h0063404.

Squire LR. 2004. "Memory Systems of the Brain: A Brief History and Current Perspective." *Neurobiology of Learning and Memory* 82(3):171-177. DOI: 10.1016/j.nlm.2004.06.005.

Stephens HM. 2014. "Performance-Based NDE Personnel Qualifications." In *ASME Pressure Vessels and Piping Conference, Volume 5: High-Pressure Technology; ASME NDE Division;*

22nd Scavuzzo Student Paper Symposium and Competition, pp. PVP2014-28427. July 20–24, 2014, Anaheim, CA. DOI 10.1115/PVP2014-28427. ASME.

Stevens A and P Coupe. 1978. "Distortions in Judged Spatial Relations." *Cognitive Psychology* 10(4):422-437. DOI: 10.1016/0010-0285(78)90006-3.

Stevens JC and HB Savin. 1962. "On the Form of Learning Curves." *Journal of the Experimental Analysis of Behavior* 5(1):15-18. DOI: 10.1901/jeab.1962.5-15.

Stevenson HW. 1954. "Latent Learning in Children." *Journal of Experimental Psychology* 47(1):17-21. DOI: 10.1037/h0060086.

Stroud JB. 1932. "Effect of Complexity of Material Upon the Form of Learning Curves." *American Journal of Psychology* 44(4):721-731. DOI: 10.2307/1414534.

Taylor JG and PC Smith. 1956. "An Investigation of the Shape of Learning Curves for Industrial Motor Tasks." *Journal of Applied Psychology* 40(3):142-149. DOI: 10.1037/h0040794.

Taylor K and D Rohrer. 2010. "The Effects of Interleaved Practice." *Applied Cognitive Psychology* 24(6):837-848. DOI: 10.1002/acp.1598.

Thomas DR and LJ Lopez. 1962. "The Effects of Delayed Testing on Generalization Slope." *Journal of Comparative and Physiological Psychology* 55(4):541-544. DOI: 10.1037/h0046103.

Thomas MS and JL McClelland. 2008. "Connectionist Models of Cognition." In *Cambridge Handbook of Computational Psychology*, ed: R Sun. NY: Cambridge University Press.

Thorndike EL. 1911. *Animal Learning*, NY: Macmillan.

Thurstone LL. 1919. "The Learning Curve Equation." *Psychological Monographs* 26(3):1-51. DOI: 10.1037/h0093187.

Tolman EC. 1948. "Cognitive Maps in Rats and Men." *Psychological Review* 55(4):189-208. DOI: 10.1037/h0061626.

Tolman EC and CH Honzik. 1930. "Introduction and Removal of Reward, and Maze Performance in Rats." *University of California Publications in Psychology* 4:257-275.

Tulving E. 1967. "The Effects of Presentation and Recall of Material in Free-Recall Learning." *Journal of Verbal Learning and Verbal Behavior* 6:175-184. DOI: 10.1016/S0022-5371(67)80092-6.

Tulving E. 1972. "Episodic and Semantic Memory." In *Organization of Memory*, eds: E Tulving and W Donaldson. Oxford, England: Academic Press.

Tulving E. 1984. "Précis of Elements of Episodic Memory." *Behavioral and Brain Sciences* 7(2):223-268. DOI: 10.1017/S0140525X0004440X.

Tulving E and DL Schacter. 1990. "Priming and Human Memory Systems." *Science* 247:301-306. DOI: 10.1126/science.2296719.

Weiner N. 1954. *The Human Use of Human Beings: Cybernetics and Society*, Boston, MA: Houghton-Mifflin.

Woodworth RS. 1938. *Experimental Psychology*, New York, NY: Henry Holt and Company.

Woodworth RS and EL Thorndike. 1901. "The Influence of Improvement in One Mental Function Upon the Efficiency of Other Functions. (I)." *Psychological Review* 8(3):247-261. DOI: 10.1037/h0074898.

Appendix B – Training and Instructional Psychology Research

Summary

This appendix reviews research from the science of training and instruction as it pertains to the core issues in nondestructive evaluation (NDE) training—amount of time spent in the field to develop skill, the progressive loss of skill over periods of disuse, and the importance of specific training and instructional factors such as practice, feedback, active and observational learning, visualization and multimedia, and simulation.

The first question addressed concerns the extent to which findings from laboratory studies of learning and memory are generalizable to application settings. Two very consistent findings in laboratory studies are (1) the superior retention of material that is studied in a spaced rather than massed fashion, and (2) the facilitative effects of knowledge testing. The research reviewed showed a number of demonstration studies conducted with medical students and illustrated that both study spacing and knowledge testing enhance retention beyond what is obtained with the same study time that is massed or simply additional study time. They key principle resulting from this research is that **distributed (spaced) practice of occupationally relevant material and knowledge testing of that same material yields superior retention to simple re-study**. The beneficial effects in these studies were obtained on the basis of planning for the presentation or testing of curriculum material in such a way as to yield spaced presentation or knowledge testing.

The review identified very few articles that specifically focused on the topic of on-the-job training (OJT). Distinctions were made between on-the-job experience and OJT. The latter is considered to be a structured process that is overseen by more senior professionals; simple "experience" in the job setting does not necessarily lead to specific knowledge acquisition and ability to apply it. In its most elaborated form, OJT instruction employs many different techniques that have been shown to be effective in teaching, such as incremental instruction, shaping of responses, feedback, etc. Evaluation in the OJT context is often linked to instruction, with gradual increases in task responsibility by the trainee and the instructor appraising performance in operational tasks. The key principle for OJT is that **OJT is a systematic process based on analysis of the task to be performed, development of instructional situations from the operational environment, employment of instructional techniques as warranted, and evaluation of performance based on increasing competence in operational tasks.** To the extent that certification or regulatory processes specify tasks or operational circumstances to be completed, as in aviation, the more likely field experience requirements are to be structured to accomplish them.

In the training and instructional psychology literature, feedback is considered in the context of the overall learning process, as distinct from the more test-item-specific feedback addressed in laboratory studies. The instructional process entails an extended set of interactions between learner and instructor and provides opportunities for feedback at multiple levels. Feedback theory has identified four levels at which learners can be provided with performance information: task, process, self-regulation, and self. The task level is addressed with feedback concerning specific performance on test items—e.g., whether the answer is correct or incorrect, or by providing the correct answer. Task feedback is more effective when it is specific, such as individual item responses, rather than general, such as overall test performance. Process feedback is concerned with what learners do in order to accomplish a task, including their

method of problem solving, strategies, concepts used, etc. This level of feedback is aimed at improving task performance by enhancing the process. Self-regulation involves the learner developing the ability to monitor and appraise their own performance; it is generated internally but is based on prior task and process-level feedback. Feedback concerning "self" tends to involve instructor praise that is not task- or process-specific, and thus not particularly effective. A key principle from the review is that **feedback is a multi-layered process addressing task performance accuracy, the process by which the task is performed (i.e., how it is performed), and can lead to internal cognitive structures for self-monitoring**.

Further work in the area of feedback addressing delayed instruction, error management, scenario reviews, and the use of supportive materials (learning criteria, checklists, reflection sheets) suggests that **feedback is a constructive process which can actively engage the learner**. This active engagement facilitates development of self-monitoring. **Feedback is thus a continuous process of constructive evaluation that can become part of ongoing operations**.

A variety of research findings indicate that as students engage with material in more active ways than being passive recipients of information, their comprehension, retention, and ability to apply knowledge improves. This pattern of findings has been described as ICAP (Interactive, Constructive, Active, Passive), to characterize learning modes with increasing levels of student engagement both with study material and with each other. Increases in activity level, for example from passive to active, may simply involve an instructional manipulation such as highlighting material and reviewing it rather than just reading. Constructive and interactive modes entail more elaborate interactions with the material, including conceptual mapping, generating self-explanations, and group discussions about material. The general principle illustrated by this work is that **as learning modes involve increasing level of material and student engagement, overall benefits in terms of knowledge structure, retention, and application are improved.** Active learning modes can involve a variety of instructional techniques, including studying worked examples and generating self-explanations. More traditional approaches such as lecture and discussion can also be adapted to incorporate higher levels of student activity.

Observation is a core element of all learning processes, whether in a classroom setting or on the job. In the context of instructional techniques, observational learning is specifically defined as watching a role model carry out target behaviors (behavior modeling training [BMT]), or parts of the job task (demonstration-based training [DBT]). In either case, the learner sees examples of expected performance and can be provided with the opportunity to reproduce what they have seen. Observational learning involving full BMT would entail watching a scripted sequence of actions, whereas DBT would involve selecting specific elements of a job to illustrate. Both approaches complement other training techniques and provide concrete examples of expected performance. The key principle for observational learning is that **student observation of whole or part performance can be flexibly implemented in training or operational settings to illustrate integration of knowledge.**

The multimedia instruction hypothesis—that learning is better with words and pictures than with words alone—has been confirmed in numerous experiments. The superior effect of multimedia is related to the structure of the human cognitive system, which encodes information in separate channels most appropriate to the sensory stream. Research has elucidated a number of design principles of visualization and multimedia that serve to focus attention through information selection, enhance essential information processing through organization, and integrate the material with other knowledge. Visualization and multimedia development capabilities are widely available, and it is important to adhere to design principles to reduce excessive material. The core principle from studies of visualization and multimedia is that **learning is enhanced by using multiple sensory modes of presentation for material that is designed to select, organize, and integrate learning content**. Studies of learning conducted with material presented by intuitive instructor design versus the same material redesigned according to multimedia principles shows that learning and retention are improved by the latter.

Simulators are widely applied in many training programs, particularly aviation. The core issue from the standpoint of personnel qualification is the extent to which a simulator can be used to substitute for training in the field, whether it be operational aircraft, clinically supervised surgeries, or supervised experience during in-service NDE. This is known as “syllabus reduction,” i.e., the percentage reduction in operational training time that might be achieved with simulation. In aviation, studies over the past 60 years have reported an extremely wide range (1%–93% operational training reduction). More recent studies suggest reductions in operational aircraft time of 25%. Federal Aviation Administration (FAA) regulations adopt a conservative stance—for pilot certification, no more than 100 of the required 1500 hours of flight time may be obtained in a simulator, with more liberal substitution for approved training programs and various ratings once a pilot is certified. Studies of surgical simulation suggest that there is positive transfer from simulation to clinical experience, with the potential to reduce training time by as much as 40%. The general principle resulting from this work is that **simulation results in positive transfer of training and can be used to reduce operational experience requirements**. The literature reviewed also suggests that determining a fixed percentage of syllabus reduction and simulator transfer effectiveness is experimentally intractable due to the many different combinations of simulator vs. operational time required. General guidance is that **for high-risk jobs, simulation should be considered an “augmentation” rather than a substitute for operational experience.**

The predominant theoretical view of developing expertise is that of deliberate practice. Studies of simple skills indicate that continuing practice yields improvement over periods of years. Theory and supporting data suggest that proficiency at professional tasks develops through a series of competency levels, with most average practitioners achieving acceptable proficiency at the level of automated skill and improving no further despite long tenure in a profession. Experts are defined as practitioners who reliably achieve superior results. When opportunities for practitioners are provided to specialize and perform many more hours of work in their domain, expertise can be developed. Specific numbers of hours necessary to achieve expertise is elusive, although popular accounts have misinterpreted research to suggest that there is a 10,000-hour minimum. While 10,000 hours is a guide, it is not a hard requirement; instead, it reflects the progression over a time period in which opportunities for practice are available. The general principle that results from this work is that **practice in high volume upon tasks that challenge current skill in association with motivation and feedback can lead to higher levels of proficiency**.

The skill retention literature illustrates that skill decay is a predictable consequence across numerous types of tasks, both laboratory learning experiments as discussed in Appendix A and diverse application areas including military, aviation, and medical domains. Continuous perceptual-motor tasks—akin to vehicle control—tend to be preserved over periods of disuse, whereas tasks involving complex cognitive and procedural execution with various decision points and contingencies tend to degrade relatively quickly, often within the first several months of disuse. Evidence suggests that skill decay is somewhat mitigated by a higher level of original learning, but it is not eliminated. Parametric studies of refresher training have not been conducted, but available evidence suggests that skill levels can be re-established with training provided during the period of disuse even through such simple means as mental rehearsal. The general principle associated with this research is that **cognitive-procedural skills decline rapidly during periods of disuse and can be mitigated by training provided during the period when skills are not being exercised in an operational environment**.

With respect to NDE training, the foregoing principles suggest several potential applications across the spectrum of initial fundamental training, field experience, and skill maintenance during down-time:

- To the extent practicable, incorporating the concepts of distributed (spaced) practice and knowledge testing during course work that conveys foundational material will enhance retention of that knowledge.
- Presentation of foundational learning material through visualization and multimedia designed according to psychological principles can enhance attention to relevant material and integration with existing knowledge.
- Use of active-learning techniques in foundational training as well as operational field experience will enhance learning and retention.
- Field experience provides the opportunity for structured OJT and should be provided to junior personnel in a way that avoids the simple accrual of time in the field. Instead, specifically structured activities that provide the opportunity to execute increasingly responsible tasks under the supervision of senior personnel is the best way to implement field experience. OJT should incorporate behavior modeling and DBT as appropriate to the specific circumstances.
- Feedback across both the training and operational experiences should be routinely provided both in terms of specific task outcomes and the processes used by individuals. Active engagement of feedback can help to develop self-monitoring mechanisms and use of post-inspection debriefings can enhance the inspection process both from the standpoint of the individual examiner and the utility.
- Laboratory exercises and/or ultrasonic inspection simulation will yield transfer of training and can be a learning accelerator, but a conservative approach to substitution for operational experience is suggested by practices in other industries.
- Continued practice both through operational inspections and use of practice samples will improve skill levels. Expertise requires many years of experience to develop, particularly in circumstances where difficult cases are rare. This suggests that development of specific practice samples and scenarios that incorporate the characteristics of difficult cases that have been encountered would be a useful addition to the overall training and practice experience of examiners.
- Cognitive-procedural skill will decay during periods of disuse, and training and practice during these down-times is warranted to maintain proficiency.

These suggestions for application of principles derived from the instructional and training literature can provide impetus for review of existing NDE training and field experience practices and potentially identify opportunities for implementation.

The review of scientific literature documented in this appendix was conducted to determine the extent to which there exists a basis for reducing the number of experience hours to attain an initial Level II rating and to identify material pertinent to skill maintenance via annual practice. The industry-proposed change in field experience requirements reduces the number of hours from 800 to 400, with 320 of those hours permissible as laboratory training. The overall reduction of in-field experience is thus as high as 90%, with lab or simulated flaw training substitution of 80%. No change to annual practice requirements of 8 hours is proposed.

The key findings from the literature review indicate that **there is no technical basis for reducing field experience time requirements**; greater amounts of learning time lead to greater levels of knowledge and retention. Knowledge and procedural skill decay very quickly during periods of disuse, and the rate of decay is similar for experts and non-experts. Knowledge decays from the overall level attained, so there will be lower amounts of loss for those individuals with higher levels of knowledge; complex procedures that are memory-intensive and involve multiple branching steps are particularly prone to skill decay. Studies of simulator substitution and the regulations governing the airline industry suggest that simulator training is acceptable for small substitution percentages during initial training and higher percentages for specific aircraft types and certifications (e.g., instrumented flight) once the primary certification is obtained. As an example, initial commercial pilot certification requires 1500 hours of operational aircraft flight experience, for which 100 hours in an approved simulator may be substituted. In medicine, simulation is required in various surgical curricula but does not substitute for clinical experience.

There are certain practical considerations associated with initial field experience and annual practice that warrant consideration. For new personnel with little to no experience in the field of NDE, the 800-hour requirement is warranted—this will provide the best opportunity to gain supervised experience in the operational environment. Higher amounts of field experience would pose a significant entry barrier, while lower amounts of laboratory practice reduce contextual factors that are important for skill development. **For NDE examiners who already have considerable experience in other special industry sectors, it may be appropriate to reduce the 800-hour requirement and substitute examination of practice samples to build knowledge of special nuclear industry materials and flaw characteristics. This would be coupled with appropriate Performance Demonstration Initiative qualification**.

With respect to practice and experience once initial knowledge is established, the overriding conclusion is similar to that of initial learning—more practice and experience leads to greater levels of expertise. Retraining after periods of time off has not been examined parametrically in the literature except in artificial laboratory verbal learning studies. **It is thus not possible to determine whether the currently required 8 hours of practice on operationally-similar specimens is sufficient.** Laboratory data suggest that less time is required to regain mastery after skill or knowledge decays (there is a “savings” in re-learning), but the extent to which such savings translate to operational environments is unclear. Certain skilled professions such as medicine require continuing education annually of 40 hours or more; for pilots who have not flown for more than 90 days, three takeoffs and landings in an approved visual simulator must be performed. Studies of cardiopulmonary resuscitation skill suggest that continuing training and feedback embedded into the operational environment on a case-by-case basis is the best approach to skill maintenance. Because testing has been shown to enhance retention of existing knowledge, it may be appropriate to develop further guidance on how to structure and test examiner practice on relevant samples, as there is currently no structure specified in the Code. The apparently self-directed nature of annual practice is not optimal from the standpoint of self-monitoring of knowledge; the literature suggests that people are not particularly good at judging what they should study and how much they should study. Thus, some **more specific structure and guidance of annual practice that includes studying material pertinent to upcoming examination and knowledge testing of what is studied would be appropriate.**

B.1 Introduction and Background

This appendix of the literature review is concerned with the applied science of training and instructional research. The material in Appendix A focused on laboratory and basic research studies of human learning and memory; this research is essential for establishing fundamental principles that are thought to underlie the more practical domains of job training and classroom learning. Indeed, it was the hope of mid-20th century learning theorists that the science of learning and memory would be applied in more realistic settings fairly quickly (Melton 1959). It has taken longer for the translational science process than anticipated, but there is now a clear linkage between laboratory and applied settings, as suggested by some of the examples provided in Appendix A.

While we are addressing training and instructional psychology research in the same appendix, there are distinctions between the two areas that need to be mentioned. “Training” is generally considered to be job-oriented, i.e., students—whether children or adults—learn specific knowledge and skills in order to accomplish a task. “Instruction” is a broader term and involves manipulation of the learning environment to result in experience by the learner (Mayer 2011). This encompasses methods of content delivery that include traditional classroom teaching, interpersonal elements such as feedback and praise, as well as other approaches such as computer-based or multimedia instruction. It is clear from this definition that training for job skills is based upon instruction. Instructional research has developed with a focus on *education*, i.e., the learning of fundamental knowledge and skill as preparation for more focused application in a career. A further distinction concerns criterion measures of performance—training is concerned with acceptable job performance; education uses the criterion measure of grades as a basis for breadth and depth of knowledge.

Several decades of reviews of training research have successively elaborated the multivariate nature of the domain and identified specific trends. The first comprehensive review was made by Campbell (1971) who pointed out the faddishness of the literature, and found it difficult to derive any broad general principles that could be applied across diverse subject matter. A subsequent review by Goldstein (1980) emphasized that training occurred within a systems context, generally a work organization, and that effective design, delivery, and evaluation needs to address the whole system. Cannon-Bowers et al. (1995) outlined a comprehensive model of training effectiveness, which includes the following general variables:

- Individual characteristics such as abilities and attitudes
- Organizational and situational characteristics such as specific jobs, tasks, policies, and procedures
- Training methods including structure and means of delivery
- Motivation of the trainees.

This model is very similar to the organizational systems model used as a basis for human factors task analysis of manual ultrasonic testing (UT) (Sanquist et al. 2018). Over time, these reviews have illustrated general research trends, including greater emphasis on non-traditional delivery methods such as distance learning, simulation and gaming, and team training (Salas and Cannon-Bowers 2001; Bell et al. 2017). Bell et al. (2017) reviewed 100 years of training research published in the *Journal of Applied Psychology*; of particular note for the present work was the emphasis on training design and delivery, and the following general conclusions (p. 316):

- "The effectiveness of a particular training delivery method depends on the skills or tasks being trained and the desired learning outcomes.
- "Training designs that encourage and support more active, self-regulated learning can facilitate the development of complex skills and adaptive performance.
- "The effectiveness of technology-based training is determined not by the media but by the design of the instruction and the support provided to learners.
- "Team training interventions, such as cross-training, leader briefings, and after-event reviews, can be effective tools for improving team functioning."

It is evident from these general conclusions that there is not a prescriptive heuristic that can be used to match training approaches and technologies to specific situations. Instead, analysis of the tasks to be trained, in conjunction with the ability of specific instructional approaches and technology to support training of those tasks, is the current strategy.

Training research has also followed larger societal trends that manifest in the workplace. Prior to the industrial revolution, training was done primarily by more experienced people (masters) for various crafts and skills. With industrialization, as well as the needs of global warfare, training was required for masses of individuals on many repetitive types of tasks. During this era of scientific management, organizational training functions were established with specialists devoted to developing and delivering training programs. Other social trends, including human relations and participative management, have reinforced and expanded the professionalization of training. Most recently, the concept of strategic learning has dominated organizational training. This movement entails the concept of continuous improvement, knowledge management, and a focus on the subjective experience of the trainees so that training has contextual relevance (Kraiger and Ford 2007).

Instructional psychology research is the branch of educational psychology concerned with the design of effective conditions of instruction (Gagne and Rohwer Jr. 1969). Research in this area is distinguished from other areas of educational psychology including the study of learning (an outcome of instruction), assessment (measurement of the outcome of instruction), and the study of other intervening variables such as motivation and individual differences. The principal goal of instructional psychology research is the discovery of evidence-based principles and practices that can be applied—either in traditional educational environments or in job-oriented training.

Early examples of instructional research were described in Appendix A. These include the classroom studies of recitation and testing by Gates (1917) and Spitzer (1939) illustrating enhanced retention of tested material versus re-study. In the first review of instructional psychology, Gagne and Rohwer Jr. (1969) focused upon studies that could be categorized into the eight events of instruction, including gaining and maintaining attention, establishing the conditions of learning by providing directions, presenting the material by various means, and providing feedback. Because it is not feasible in applied research to individually analyze separate instructional events, this cumbersome structure soon gave way to more focused theoretical and empirical analyses of specific situations and material.

As with training research, instructional psychology tended toward a degree of faddishness, particularly regarding technology. For a time in the 1960s–1980s, there was considerable focus on comparative effectiveness analysis of presenting the same material in different ways or by different technical means. The conclusions after several decades of research of this type is that learning is generally equivalent across technological approaches, and that head-to-head comparisons of technologies were providing diminishing returns (Pintrich et al. 1986). More recently, studies of multimedia approaches to enhancing comprehension of complex topics gave new impetus to studies of instructional technologies (e.g., Mayer 2001).

The current form of instructional research is best exemplified by examining the table of contents of the most recent *Handbook of Research on Learning and Instruction* (Mayer and Alexander 2017). This handbook divides the research into two general domains—*subject matter content* (reading, writing, history, science, mathematics, etc.), and *instructional process* (feedback, examples, self-explanations, peer interactions, learner inquiry, visualizations, etc.). These latter topics are the most relevant to the present concern of what is effective for nondestructive examination/evaluation (NDE) training, and so will receive appropriate focus in the following pages.

This appendix of the literature review will focus on material from the training and instructional research domains that are most germane to the overarching questions of learning time to mastery, skill maintenance, and refresher training. The specific major areas to be addressed include training/instruction design and delivery, expertise and deliberate practice, and skill retention. These concepts are appropriate for organizing a diverse body of material from both training and instructional research domains, cover both product—the content and process—the specific procedures employed to deliver the content, as well as the outcome, i.e., retention and transfer.

B.2 Scope of the Review

Because the training research literature is extremely voluminous, it is important to define the scope of the review in such a way that only pertinent material is included. Thus, a criterion for assessing the relevance of various bodies of research will be the potential applicability to what we understand the current practices in NDE training to be. These include such approaches as community college courses, private training schools, online training, and field experience gained by supervised inservice inspection. This will necessarily exclude material from instructional psychology that is concerned with developmental issues, such as core reading and writing skills, and concept formation in basic areas such as math and science. Conversely, there is substantial material concerning such issues as feedback theory, use of technology, behavioral modeling, and self-regulation (metacognition) that will likely have applicability. Additionally, elements of instructional psychology research, such as learning through examples, inquiring, self-explanation, etc., have analogs in applied studies of training. Although the translation is not direct, there are sufficient parallels to warrant discussion of these topics when appropriate. The primary lens through which we will view the training and instructional research literature is that of complex technical skill training; to the extent that research areas appear to have potential utility in the NDE training approaches described above, they will be addressed. General examples of such potential utility may include how core technical material is organized, presented, repeated, and tested; guidance for supervisors for providing feedback to trainees; and trainee strategies for observing and modifying their own job performance.

B.3 Training/Instruction Design and Delivery

This section will review a number of separate areas that can be considered aspects of training and instruction design and delivery. We focus on specific variables associated with how content is organized and presented—manipulations of the learning environment—and the impacts of these variables upon learning outcomes. The discussion does not address what has come to be known as *Instructional Systems Design*, a general and systematic process for training needs analysis and the design of specific course content; such material is well-covered in many standard reference texts (e.g., Gagne et al. 1992). A diverse body of material is covered in this section, from targeted demonstrations of principles defined in laboratory research, to applications of feedback theory and use of simulation and multimedia.

B.3.1 Demonstrations of Laboratory Findings

The research discussed in Appendix A included studies of the spacing effect in learning and memory, and the impact of knowledge testing on retention. The spacing effect entails superior retention of information when it is learned in sessions that are distributed over time, rather than massed. Knowledge testing has been shown to increase retention simply on the basis of the test—the attempts at recall for a test seem to enhance the ability to later retrieve information, independent of additional study time. In this section, we describe illustrative examples of how these fundamental principles of learning have been demonstrated in applied educational settings.

B.3.1.1 Spacing of Learning and Practice

Applications studies of the spacing effect have focused on classroom studies that present material in a single session, or spaced out over several days, followed by a test of retention. Carpenter (2017) reviewed a number of these studies, and the findings suggest that spacing is effective for children and adults, as well as children with specific language impairments. Spacing effects result in enhanced retention for same language vocabulary, different language vocabulary, rules of grammar, scientific concepts (e.g., food chains), history, and meteorology. The quantitative impact of the spacing effect varies across experiments, but in all cases spacing yields superior performance in tests of delayed retention—generally on the order of a 10% advantage, although in some cases the advantage can be as high as 20% (Bahrick et al. 1993).

Medical students represent a unique research cohort, in that they have predictable curriculum content that is spread out over time, and consists of scientific and applied clinical knowledge, as well as the development of skills such as diagnosis and surgical technique. A number of studies have been performed in the past 20 years assessing the application of distributed practice with medical students.

Kerfoot et al. (2007) provided the first demonstration of the utility of spaced learning and study for specialized clinical knowledge necessary to pass board-qualifying examinations in urology. This study compared the performance of a group of students who received a single examination packet with 96 study/review questions and explanations, with groups of students who received daily emails of 1 or 2 questions and review material, over a 24-week period in which material was repeated at various intervals. A subset of 32 items were tested for long-term retention at various delays following completion of the study period. Results are shown in Figure B.1, which compares the massed practice (bolus) group, with the spaced practice groups.

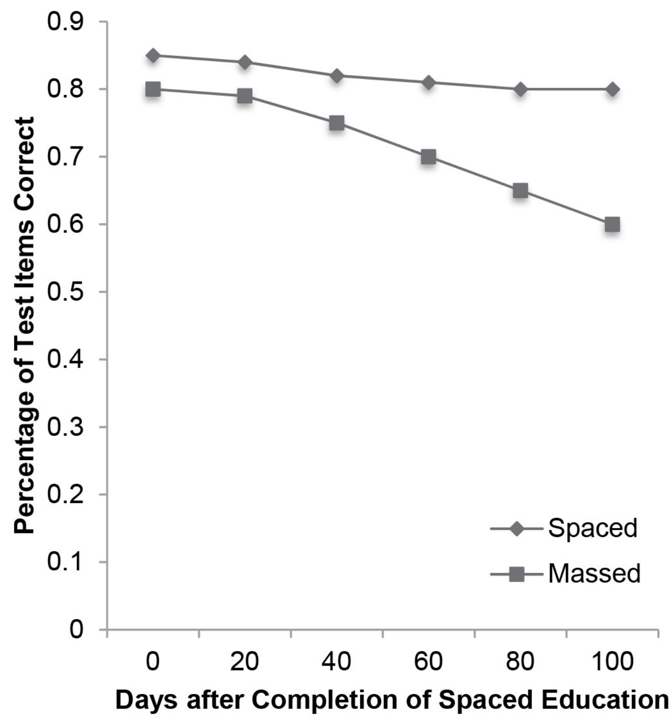

Figure B.1. Scores on 32 item practice test for massed and spaced study groups (based on (Kerfoot et al. 2007).

There is a significant effect of spaced practice, which includes stable maintenance of knowledge over time, whereas the massed study group shows lower overall performance and a decline in knowledge over time. This performance advantage was maintained at modest levels 2 years after the initial study period using different test items that were different from the original test, but derived from the same pool of study material (Kerfoot 2009). Subsequent studies by the Kerfoot research group have shown that online spaced education is associated with improvements in diagnostic skills on test items that have not been seen before, as shown by performance on test histopathology samples, and in decreasing the frequency of inappropriate prostate screening tests ordered in actual clinical practice (Kerfoot et al. 2010a; Kerfoot et al. 2010b). Figure B.2 illustrates the clinical impact of the spaced education program.

The spacing of practice also has a beneficial effect upon the psychomotor skills necessary for performing surgery. Studies of component skill training using surgical simulators have demonstrated that single-session massed practice is inferior to practice that is distributed across multiple periods in a single day (Mackay et al. 2002) or over several days that are either narrowly or widely spaced (van Dongen et al. 2011). The beneficial effects of spaced training are evident in both immediate and longer-term (2 months to 1 year) tests of retention (Spruit et al. 2015). There is also evidence that surgical skill training that involves *interleaving* of observations of correct and incorrect performance, coupled with manual practice, is superior to massed observations of either type of performance (Welsher and Grierson 2017).

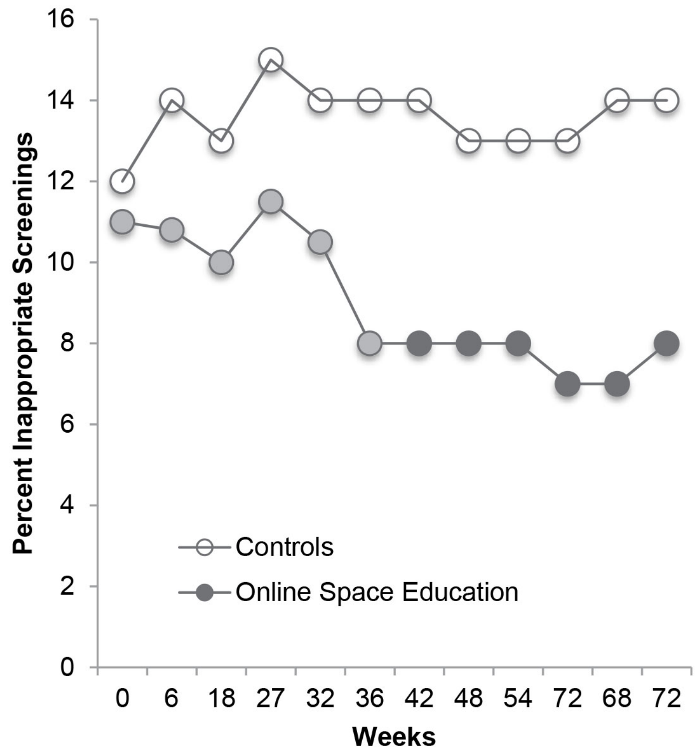

Figure B.2. Reduction of inappropriate prostate screening exams ordered by physicians in online spaced education and controls. Gray filled circles represent training period, black filled circles occur following the training period (based on Kerfoot et al. 2010b).

The extent to which distributed practice is beneficial to actual surgical performance was assessed by Moulton et al. (2006). This study employed surgical residents who were trained with video and practice sessions to perform microvascular operations. The massed group received all training in a single day, and the distributed group received the same training material in one session per week for four weeks. The practice aspects of this study entailed connecting PVC artery models, and turkey thigh arteries, under the supervision of experienced microsurgeons. Outcome testing involved performance of similar procedures on test material, and upon anesthetized rats, in which expert evaluations of performance were made (experts were blinded to the training groups). The results indicated that on a retention test one month after training, the distributed group showed better performance on measures of time, number of hand movements, and expert global ratings); similarly, on the transfer test of live rat surgery, the distributed group out-performed the massed group on most expert rating measures. There was also a non-significant trend indicating better clinical outcome for the distributed training group.

B.3.1.2 Knowledge Testing

Application studies of knowledge testing have shown that the beneficial impact seen in the learning laboratory can be extended to standard educational settings. Initial reviews of classroom-based testing (Bangert-Drowns et al. 1991) focused on the frequency of testing using traditional paper and pencil methods, and indicated that student performance on final criterion tests can be enhanced by giving at least one test during the course of instruction. The positive impact of quizzing during a semester-long course has been shown to apply to test questions that are different than the quiz questions, suggesting that generalization has occurred.

More recent work has focused on computerized quizzing using audience response systems, otherwise known as “clickers.” Clickers are handheld devices that are wirelessly connected to a classroom instruction computer system, thereby permitting instructors to pose questions throughout a lecture and to tabulate responses immediately (Caldwell 2007). While these systems are generally used to engage students in large-scale lecture classes (e.g., several hundred students), the basic effects of “on-the-fly” quizzing are applicable in much smaller classes and with peer interaction instruction (Caldwell 2007). Mayer et al. (2009) reported a study in which students responded to 2–4 questions per lecture via clicker, via paper-based quizzing, or no in-lecture questioning. The results showed that the students using the clicker-based question-response approach performed 1/3 of a grade point better than the other two groups, which were equivalent. Results of a variety of studies such as these are interpreted to suggest that quizzing engages students in a more active, generative approach to learning.

A different approach to knowledge testing was demonstrated by Butler et al. (2014), which incorporated homework-based questions that used space repetition of concepts and correct answer feedback. While this technique did not use in-lecture testing with immediate scoring, there was substantial structure in the homework assignments. Students using these approaches also showed better performance than controls. The results of Butler et al. (2014) suggest that knowledge testing can be implemented in alternative ways, which may be beneficial for de-centralized learning programs.

The potential advantages of test-enhanced learning in medical education is discussed by Larsen et al. (2008). They point out that clinical expertise and problem solving are heavily dependent on a base of clinical knowledge, and that test-enhanced learning has been shown to enhance memory for such information. However, other studies suggest that despite requirements for fundamental topic coverage in resident medical education, there is no difference in long-term knowledge retention between residents who attend didactic conferences and those who do not (Larsen et al. 2009). These factors present a challenge to medical education researchers to address the prospect of testing as a learning tool, in addition to its traditional role in assessment. Several subsequent research studies have investigated this issue.

The effect of repeated study of medical material versus repeated testing was investigated by Larsen et al. (2009). This experiment involved two groups of medical residents who were initially taught two different medical conditions in a standard lunchtime didactic conference. Over the next six months, one group of students studied the teaching material on repeated occasions, while another group took repeated tests. The 6-month follow-up knowledge test showed superior performance for the repeated testing groups. This large effect of repeated testing was also associated with positive ratings from the subjects, who indicated their willingness to participate in similar repeated testing for other knowledge topics. Similar findings were reported by Larsen et al. (2013b) in a study of combining test-enhanced learning with self-generated explanations of correct answers. This study compared performance on topical subjects taught to medical students in the following conditions: study-only, study with self-generated explanations, testing without self-explanations, and testing with self-explanations. The results showed the expected superiority of material learned under conditions of repeated testing (a 16% advantage), and a smaller effect of self-explanation (9% advantage in study-only conditions, 4% advantage in testing conditions).

In addition to enhanced knowledge retention on conventional assessment tests, there is evidence that testing as a learning tool facilitates knowledge transfer and clinical application. A simple demonstration of this effect was reported by Baghdady et al. (2014) using dental radiographs as study material. The experiment compared two groups, one of which was shown radiographs of four specific conditions and explanatory learning material, followed by a test of basic science knowledge underlying the radiographic presentation of the pathological condition. The other group was simply shown the radiographs and learning material and provided with additional study time after the learning session. A follow-up test one week later involved presentation of 22 oral radiographs for which the subjects were to select the correct diagnosis from the four pathological conditions originally studied. The group that learned the material followed by a test of knowledge showed superior performance on the diagnostic test (9% advantage on test immediately following learning and a 4% advantage one week later).

Larsen et al. (2013b) reported a study of test-enhanced medical learning comparing study-only conditions with learning using repeated written testing and learning using interactive standardized patient scripts. These scripts use actors presenting complaints relevant to the specific medical condition being learned. Final knowledge testing used written test items and standard patient scripts that addressed the clinical condition of interest but differed from the scripts used during the learning phase. The key finding from this study was the superiority of testing over repeated study, and the superiority of standardized patient testing as a learning tool over written testing (Figure B.3).

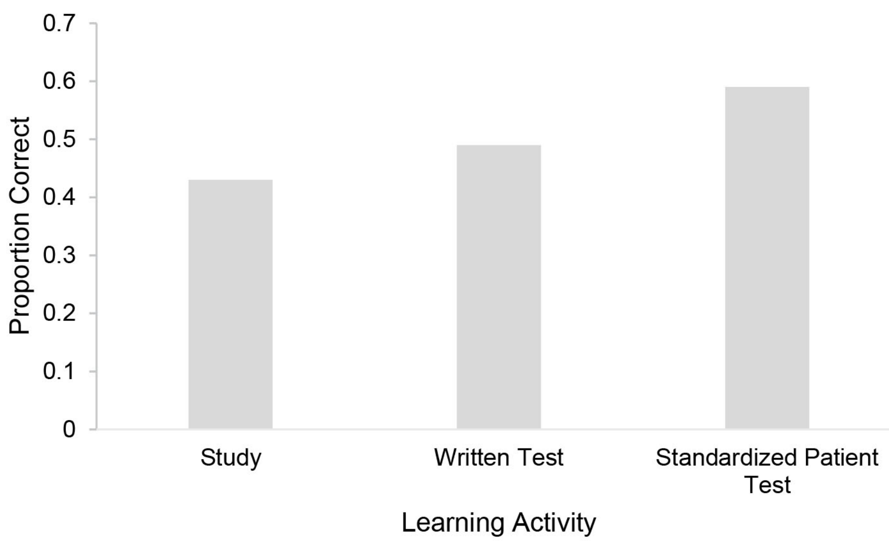

Figure B.3. Mean proportion correct on a final test of medical knowledge using standardized patient based on (based on Larsen et al. 2013a).

The Larsen et al. (2013a) results show the expected superiority of testing as a study tool, but more importantly provides a contrast across types of test. In this case, standardized patients—a frequently used medical simulation tool—were used to test knowledge gained during initial learning. Follow-up questionnaires with the subjects suggested that the standardized testing caused many of them to mentally step through the logic of diagnosing specific clinical conditions, which is similar to procedural reinstatement in transfer of training.

B.3.1.3 Translational Science Issues

The studies reviewed in this section illustrate that basic findings from the learning and memory laboratory can be implemented in a variety of educational contexts, including classroom learning of simple and more complex materials, as well as individualized instruction of complex medical knowledge. The performance improvements seen with spaced practice and knowledge testing appear to be quite general and adaptable to a variety of circumstances. These clear benefits unfortunately appear to be limited to a relatively few demonstration studies, with little progress toward implementation more broadly across the educational system. Roediger (2013) discusses the numerous institutional barriers to implementation, including entrenched philosophies of education that survive despite evidence to the contrary, as well as faddish and economically motivated technology and instructional programs lacking an evidence base. It is also worth noting that considerable “front-end” work needs to be done to modify existing curricula to incorporate the simple mechanisms of spacing and knowledge testing. This would include textbook modification, development of alternative presentation material, more frequent testing, and potentially, changing standards of assessment to address improved baseline performance. While Butler et al. (2014) assert that translating the basic cognitive science knowledge to broad classroom application does not require a Herculean effort—at least at the individual course level—institutional change is more unwieldy. The results of the studies of medical education reviewed above provide some encouragement, in that training medical professionals is a key element of our knowledge infrastructure, and one that despite institutional inertia, is oriented toward experimentation and translation of results to action.

B.3.1.4 Relevance for NDE Training

The studies reviewed above illustrate that the basic phenomena of the spacing effect and knowledge testing can enhance learning and retention across diverse instructional settings and material. Because NDE entails basic learning in various areas of physics and engineering, as well as application of this knowledge to evaluating and diagnosing material flaws, there would appear to be an opportunity to conduct small-scale implementations to determine feasibility. One of the principal challenges in doing this was mentioned above—the tendency toward institutional inertia and the need for extra effort to modify existing training material and practices.

While community colleges and private training organizations have little incentive to change current practices, the desire of the American Society of Mechanical Engineers (ASME) Code Committee to reduce field experience hours may provide an impetus toward experimentation. For example, the currently required 800 hours of field experience may potentially be reduced if it were shown that spaced repetition of practice sample exams yielded a level of procedural knowledge that is acceptable with less than 800 hours of service time. Such a demonstration experiment would need to be coordinated by various concerned entities (utilities, Electric Power Research Institute [EPRI], vendors, ASME) and overseen by a regulatory body (e.g., the U.S. Nuclear Regulatory Commission [NRC]).

Perhaps the most immediately applicable research reviewed above is the work of Larsen et al. (2013b), in which students initially acquired clinical knowledge from standard instruction, which was then reinforced through testing with standardized patients. The standardized patients simulated what would actually occur in a real-life clinical consultation, and therefore required not only recollection of factual knowledge, but use of that knowledge to guide questioning and assessment during simulated patient interactions. This type of learning suggests a potential application in NDE training, i.e., initial learning about metal properties, stresses, and geometric reflectors with standard instruction, followed by reinforcement of that knowledge with simulated inspections. This might initially be attempted using low-fidelity approaches, such as written descriptions of the components to be inspected, and their operating conditions, and requiring students to verbalize their process, with an expert instructor revealing more information as the “inspection” progresses. This would also permit the opportunity for feedback and correction as necessary.

B.3.2 On-the-Job Training

The experience hours required for various levels of nondestructive testing certification are specified in ASME Code, and generally entail a specific number of hours of field experience in the method of examination for which the candidate is seeking certification. Level II ultrasonic certification, for example, currently requires 800 hours of field experience. This “time in method” requirement is intended to involve supervised performance of setting up and executing test procedures, observing, calibrating and other activities specific to the technical aspects of the examination. This process is basically a form of on-the-job training (OJT) meant to develop requisite technical and procedural knowledge by means of supervised job performance and is distinct from other organized training programs. It is noteworthy that Level III NDE personnel have expressed concerns that the way more junior staff are accumulating field experience hours is suboptimal, in that making thickness measurements does not provide the more relevant experience of a full component examination (Sanquist et al. 2018) and that opportunities for such experience are diminishing.

The general rationale of accruing experience hours under the supervision of senior personnel is applied across many professional settings, including aircraft piloting, surgery, and psychotherapy as just a few examples. The process is meant to serve a quality control function by ensuring that only appropriately trained and experienced individuals enter a profession (Cathcart and Graff 1978; Adams et al. 2004). It is also the case that professional certification or licensing examinations must be preceded by the stipulated experience hours, i.e., demonstrated experience hours are an entry requirement for examination pursuant to licensing.

In the case of ultrasonic NDE, employers certify the experience hours and often assume the responsibility of sending candidates for Performance Demonstration Initiative (PDI) qualification testing. Thus, it is in the interest of both candidate and employer to ensure an adequate level of preparation for the testing process, generally through supervised practical experience. Employer “written practices” for certifying acceptable experience hours are documented and form the basis for quality control per ASME Code. At the present time, there is no uniformly accepted standard for the content of the required experience hours—simply a quantitative statement. This can lead to substantial variations in how experience hours are obtained, as mentioned above, and may result in unprepared candidates sitting for PDI examinations. Thus, OJT as currently implemented within the nuclear ultrasonic examination sector may warrant enhancement by means of a more structured approach to the supervised experience hours.

There are few research studies of OJT to guide evaluation of how such training is done or how it could be improved. In general, OJT should be distinguished from on-the-job *experience*. The latter may simply involve exposure, while OJT specifically involves supervision and feedback by more senior personnel as work is performed. According to Bailey (1993), “simple exposure to the job is not likely to achieve the expected results….programs that simply place young people on the job to gain work experience are not effective.” More generally, there is a prevailing opinion that workplace training is unstructured and relatively ineffective, and that little attention has been devoted to how on-the-job learning experiences can be enhanced and evaluated. Despite the shortcomings of the applied research literature, there is a base of knowledge from instructional development and tutoring research that can be used to establish a framework for more effective OJT.

The prescriptive model of OJT described by Semb et al. (2000) is based on three principal components: (1) assessment of what needs to be taught, (2) instruction in the job elements identified in the assessment phase, and (3) evaluation of the candidate performance throughout the OJT process. These are the basic elements of what is more generally considered a systematic approach to instruction.

B.3.2.1 Assessment

Assessment of what needs to be taught is most systematic if it is based on a task analysis of the job to be performed. Prior work by the NRC and Pacific Northwest National Laboratory (PNNL) (Sanquist et al. 2018) and the ASME NDE (ANDE) group have identified the general elements that should be addressed in ultrasonic NDE OJT. The following task elements were listed in an EPRI (2008) report addressing the potential for “fast-track” training of nuclear NDE examiners:

- Review geometric descriptions and history of the component to be examined
- Select appropriate examination techniques for the job
- Prepare and calibrate the inspection system for the job
- Scan the surface area to detect and characterize flaws in the material
- Discriminate flaw indications from other signals
- Record and report examination results.

These high-level tasks comprise the core of what needs to be trained; additional lower-level tasks, such as equipment assembly and signal interpretation, are encompassed in the documented task analyses.

In addition to the overarching OJT content specified by task analysis, there is an individualized aspect to instructor-trainee assessment—basically determining what the trainee knows in relation to the task analysis areas. This is generally accomplished through a question-demonstration process. For example, on the procedural task of UT equipment assembly, the supervisor would ask “do you know how to assemble the equipment?” If the answer is “yes,” this would be followed by demonstration, with potential questions along the way addressing such procedurally-specific issues as calibration, test block temperature, etc. The following elements described by Semb et al. (2000) are prescriptive steps for the OJT assessment phase:

- Analyze content and tasks of instructional program to determine what needs to be taught or use available job specifications
- Diagnose prior knowledge by asking trainees to explain what they know
- Diagnose prior skills by observing trainees’ performance on tasks before instruction
- Assess transfer abilities by questioning and observing on novel examples.

B.3.2.2 Instruction

The instruction phase of OJT has been compared to tutoring or coaching, within the constraints of an operational environment (Semb et al. 2000). While the task analysis can set the parameters for what needs to be taught, the realities of operational inspection may preclude performing certain types of tasks or procedures. In the extreme case, utilities may not accept trainees as part of the inspection cost proposals made by vendors, thus severely restricting learning opportunities (EPRI 2008). However, assuming that some operational instruction opportunities are available, the tutoring literature suggests the following practices that may be employed by OJT supervisors, within the constraints of an operating environment:

- Incremental instruction
- Prompting
- Practice
- Feedback
- Examples
- Systematic questioning and explanation building
- Demonstration and modeling.

All of these approaches tend to combine in OJT for procedural tasks, particularly demonstration and modeling, if the trainee is a complete novice. This might be followed by practice, feedback, and other techniques as appropriate to the situation. OJT should entail opportunities for trainees to demonstrate transfer—application of what they have learned to novel situations. Table B.1 (from Semb et al. 2000) delineates a number of OJT strategies and references to supporting research.

Table B.1. OJT Strategies (from Semb et al. 2000)

OJT Concept	Definition
Incremental instruction	Divide material into sequence of small steps (Keller 1968, 1974; Broadwell 1986)
Shaping and fading	Using process of progressively complex steps, trainee's behavior is contingently reinforced or corrected by supervisor prompts, which are gradually decreased as trainee's performance approaches desired state (Skinner 1958)
Scaffolding	Level of trainer's intervention should vary with ability of trainee (Wood et al. 1976) Metaphor from building construction to describe ideal role of trainer (Greenfield 1984) "Advance organizers" – use of higher-order terms to aid conceptual understanding of new material (Ausubel 1960) Use of "procedural prompts" or "signal words" to assist trainee when needed (Rosenshine and Meister 1994)
Modeling	By trainer first (Broadwell 1986; Knox 1986); by trainee first, if safety factors allow (Knox 1986)
Practice	Trainee explains or demonstrates several times (Broadwell 1986; Knox 1986)
Opportunity to respond	Extent to which instructional stimuli elicit trainee's behavior for shaping (Greenwood et al. 1984)
Questioning	Five-step frame (Graesser 1993) Recall is better when material is actively organized by the trainee (Brown 1975; Palinscar and Brown 1984) Semantic elaboration or "deep" processing enhances memory (Craik and Tulving 1975)

Many of the elements in Table B.1 are self-explanatory. Two of the concepts—opportunity to respond and the five-step frame of questioning—need further definition. The opportunity of the student to respond is essentially the amount of time or number of occasions in which the trainee has a chance to demonstrate their knowledge. Studies of traditional classrooms have shown that about 25% of the time, students are engaged in oral reading, responding to questions, and writing, with another 50% devoted to passive listening or watching. Increases in active student engagement, through use of peer tutoring, has been shown to improve achievement.

The five-step frame for questioning was described by Graesser (1993) on the basis of naturalistic observations of tutors and students in a variety of contexts. A pervasive pattern was observed consisting of the following elements: (1) tutor asks a question, (2) student answers, (3) tutor gives short feedback on answer quality, (4) tutor and student collaboratively improve on answer quality, and (5) tutor assesses the student's understanding of the answer. It was generally found that tutor questions and feedback were driven by scripts based on the knowledge to be imparted, and that there was little time devoted to diagnosing errors in responding and correcting flawed knowledge that led to the error.

B.3.2.3 Evaluation

Evaluation of trainee performance is the third element of OJT. This is often intertwined with instruction, as the supervisor will correct errors in the trainee's process and final performance, particularly in procedural tasks. Because of the informality of many OJT situations, the point at which evaluation becomes distinct from instruction is often unclear, but a meaningful break point is toward the end of the period specified for hours of experience. It is in the interest of NDE vendors to know that their trainees are ready to undertake the next step of PDI qualification.

Evaluation of *process* entails the sequential steps taken by the trainee to accomplish a task—are all of the steps done? Are they done in the right or preferred order? Examples from NDE might include the steps taken to characterize and scan a weld—in this case, the individual steps might be performed in somewhat variable order, but they all must be accomplished in order to obtain the end result. Evaluation of the *product* of the task involves judgments of quality and/or accuracy. In the case of NDE examination, the product would be the inspection result—a determination as to whether a component did or did not show indications, and whether these indications are judged to constitute a flaw.

The general recommendation for evaluation is to use the task analysis from the assessment phase as a benchmark for what the trainee has learned. This may be broken down into more detailed sub-task components as part of a checklist for procedural steps and correspondence between the supervisor and trainer for products of outcomes.

B.3.2.4 Relevance for NDE Training

Much of the relevance of OJT to NDE training has been discussed above—the requirement for field experience hours, the lack of content specificity for those hours (this is an issue broadly across professions), and the relative lack of control over operating conditions versus the high fidelity of the training environment. The core issue for NDE examiners is the difficulty of achieving the requisite experience hours to certify for Level II—and hence sit for PDI examinations—and the extent to which those hours prepare them for actually doing PDI procedure examinations. These issues were discussed extensively in a series of reports by EPRI (e.g., 2008) in describing a concept for a "fast-track" program to expose trainees to many more flaws than they would see in 800 hours of field experience, by means of a laboratory-based program administered by selected training organizations. At the time the 2008 EPRI report was written, the concern was being able to supply sufficient numbers of examiners for inspections due to an aging workforce and anticipated new construction. The new construction issue is no longer a factor, but the aging workforce is still a consideration. It seems likely that alternatives to OJT via inservice inspection field experience will be necessary, both to generate newly trained examiners and to qualify existing Level II and IIIs for inspections in the nuclear sector. The extent to which exposure to laboratory specimens can be substituted for field experience is a problem that is amenable to analysis, but it would need to be approached from the standpoint of formative assessment and various scenarios for qualification, potentially with some limits based on how training was received.

B.3.3 Feedback

The discussion of feedback in Appendix A focused on results from memory experiments. The general finding from these studies was that feedback that is delayed rather than instantaneous, and that provides the correct answer, is superior to other forms. The controlled nature of laboratory memory experiments makes these comparisons possible. However, generalizing to the broad realm of instruction and training is difficult due to the many additional variables involved. These include the type of material studied, conditions under which learning occurs, who delivers feedback—as well as how and when, what the feedback is aimed at—task performance, performance process or the characteristics of the learner, etc.

This section discusses a range of data and theory that are applicable to the wider range of learning situations encountered in NDE training. Of particular importance for NDE training is the linkage of feedback to OJT, as described above. Supervisory oversight of trainees as they learn how to do the job will generally involve continuous feedback on a variety of job elements and should be provided in such a way as to enhance rather than impede learning. The general findings described in this section draw from Hattie et al. (2017), Hattie and Timperley (2007), and Kluger and DeNisi (1996). These researchers have integrated data from studies as diverse as laboratory studies to classroom field research and have generalized results accordingly.

B.3.3.1 Background

Feedback has been uncritically accepted as a positive influence on learning, despite the complexity of learning situation and feedback variables that can be encountered in training and instruction (Hattie et al. 2017). Kluger and DeNisi (1996) conducted a meta-analysis of feedback studies that showed while feedback has a generally moderate and positive effect on performance, a full 38% of the studies showed zero or negative effects. Further analysis of various feedback moderator variables suggested that it is important to consider the specific nature of the task and how feedback is provided in order to have a beneficial effect. For example, it was shown that praise—generally considered a good thing—had minimal effects upon performance, and that feedback that was discouraging or directed toward a physical task showed negative effects. Research investigating physical tasks show that a delay of at least several seconds is important, and that instantaneous feedback may interfere with a performer's ability to pay attention to what they are doing (Swinnen et al. 1990). Kluger and DeNisi (1996) proposed a Feedback Intervention Theory (FIT) that addresses learner attention to the task results, the process of performing the task, self-monitoring, and learner characteristics. This model has been elaborated by Hattie and Timperley (2007) and Hattie et al. (2017) and is discussed further below.

B.3.3.2 Feedback Theory and Application Principles

In the learning environment, feedback is generally considered to be an *intervention*, i.e., a manipulation that involves presenting information to the learner in some way in order to—hopefully—enhance the learning process and task performance. Feedback theory addresses the principal element of feedback—*information*—in terms of what information is provided, how it is provided, and the resulting focus of attention on the part of the learner. FIT, originally proposed by Kluger and DeNisi (1996), discussed it as a way for learners to compare their performance against a standard, and that the information provided directed attention to various aspects of the learning situation, such as the task itself, the process of performing the task, and emotions and motivation about the task or one's self. Hattie and Timperley (2007) incorporated the attention-directing aspect of feedback into a multi-level model shown in Figure B.4.

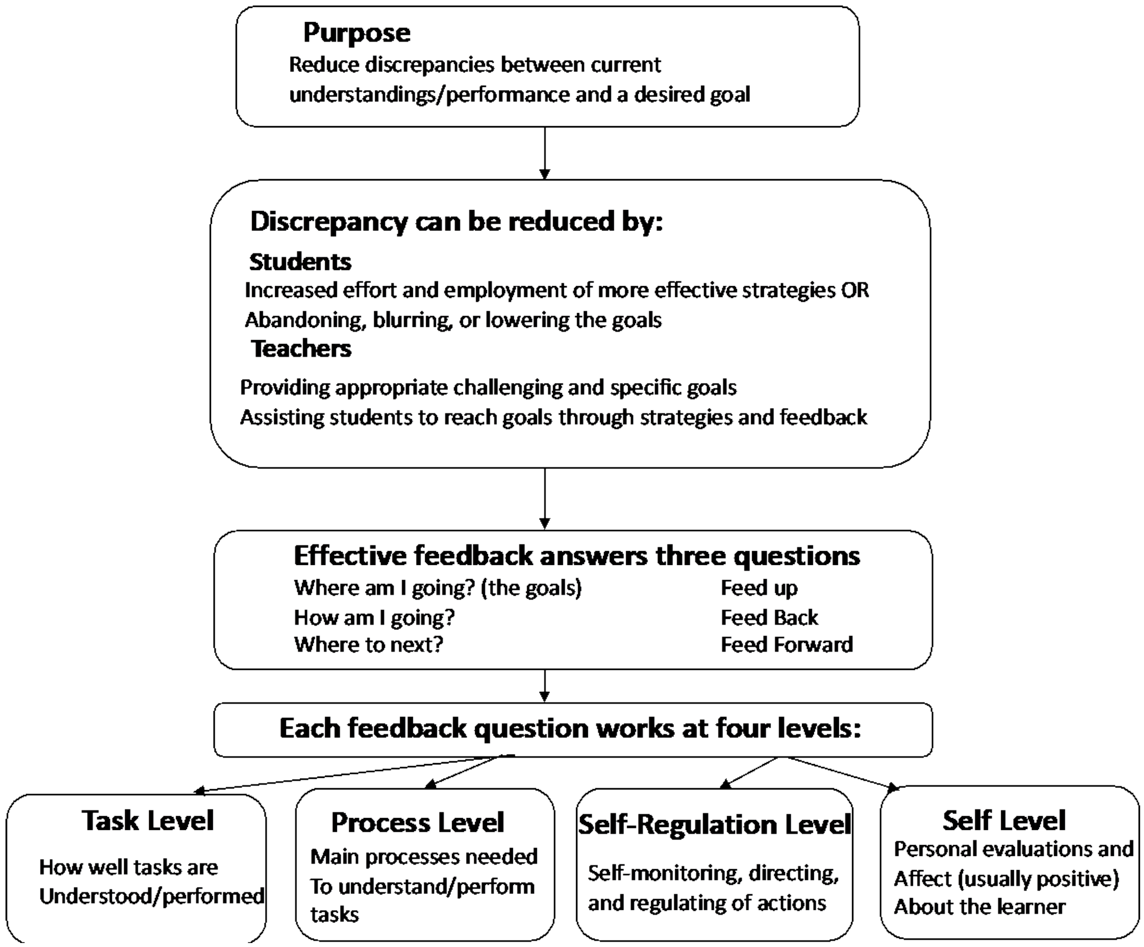

Figure B.4. Multi-level model of feedback based on Hattie and Timperley (2007).

In this multi-level model, feedback works by reducing the discrepancy between current and desired levels of performance. Performance is a multi-dimensional construct in that a *task* is performed by a person (*self*) and involves various processes to accomplish (such as effort, specific skill application, etc.), and self-regulation of behaviors involved.

The various integrative analyses referred to above have delineated the following general characteristics of effective feedback, which are described in terms of the Hattie and Timperley (2007) model:

Task feedback is concerned with whether the learner has met the performance success criteria. Specific questions addressed include:

- Is the answer correct?
- How can the answer be elaborated?
- What did the student do well?
- Where did they go wrong?
- What is the correct answer?
- What other information is needed to meet the success criterion?

Task feedback is more effective if it is specific rather than general. For example, *outcome* feedback—such as whether a person passed a course—is less effective than *component* feedback addressing performance elements—such as tests or other outputs—that contribute to the overall outcome. For example, written instructor comments tend to improve performance more so than simple grades. Task feedback is initially more effective if it is simpler (e.g., right or wrong answer) than more complex (e.g., elaborate logical reasoning followed and discussion of correct and incorrect alternatives). The timing of task feedback should be provided at a sufficient delay in order to not interfere with performance. For example, during a skilled task such as driving, the instructor should wait until an appropriate break in the flow of movement and should avoid over-specifying the task (e.g., "look in rear view mirror" rather than "move eyes to upper right and determine if any obstructions are present").

Process-level feedback is associated with helping learners understand what they are required to do in order to achieve success at a task. Specific questions addressed include:

- What strategies did the student use?
- What is the explanation for the correct answer?
- What other questions can be asked about the task?
- Are there other parts of the task that affect the current one?
- What other information is available?
- What is the learner's understanding of the concepts and knowledge related to the task?

Process issues might include reasoning that underlies performance of the task (such as how to distinguish distractor items from correct answers), where to direct visual attention in a complex operational environment, and generally developing insight into cognitive processes for error detection and correction. This level of feedback should be introduced to facilitate a deeper understanding of the learning material or tasks to be mastered. A practical example comes from driver behavior—in adapting to right-hand drive, the initial task feedback is to "keep left"—in order to avoid oncoming traffic. As more experience is gained, feedback concerning lateral lane correction techniques can be introduced, such as "correct right" in order to avoid the unconscious habit of moving to the left side of the lane.

Self-regulation feedback concerns developing the ability to appraise and monitor one's own performance. Key aspects of feedback in this regard are, first, determining what to watch for, and second, how to interpret it. In this sense, self-regulatory feedback tends to be generated internally, but also is based on external inputs such as task and process performance. The key difference is that with increasing self-regulation, learners develop mental models that can be used to generate their own task performance when they are no longer under supervision or given structured performance feedback. The questions of concern in self-regulatory feedback include:

- How can I monitor my own performance?
- How can I evaluate the information provided?
- What can account for the results observed?
- What are the common elements of my observations?
- How does this compare to other similar situations?
- Have I achieved the task goal?

Self-monitoring feedback is illustrated by the case of driving while fatigued. Many educational programs on this topic now encourage drivers to monitor their own behavior; e.g., if their driving becomes erratic, if they continually yawn or feel like they are slipping into a sleep state, they should get off the road and rest. This advice is based on developing appropriate driving task and process models from more discrete feedback during training, thus allowing drivers to self-assess.

The last level of feedback concerns personal evaluations of the individual performer, generally provided by the supervisor or teacher. It can also involve individual judgments of self-efficacy in relation to task performance. Personal feedback, such as “great job” or “well done,” typically express positive or sometimes negative evaluations about the student but have little task-related information. It has been found that such statements do very little to shape performance, due to their lack of specificity, and the emotion-laden aspect of personal evaluative statements can lead the learner to focus their attention in the wrong areas, such as trying to please the supervisor rather than perform correctly. However, if praise is provided that focuses on how the student accomplished the task (e.g., “good work applying the concept of material density”), this can lead to enhanced feelings of accomplishment that have a beneficial effect on subsequent task performance (Hattie and Timperley 2007). While praise has generally been accepted as a universal positive in the educational environment, the effects are quite complicated and interact with student age, ability level, and motivation, and has not been shown to improved task performance unless it is task-focused. Thus, when praise is offered in a training setting it should be “specific, sincere, accurate, and earned” (Hattie et al. 2017).

B.3.3.3 Elaborations of Feedback: Delayed Instruction, Error Management, and Active Engagement

Delayed instruction occurs when students are given the opportunity to explore problem-solving approaches prior to training in specific procedures. Instead, they use prior knowledge and concepts in a domain, e.g., mathematics, to attempt solutions of problem types that have not yet been covered in lessons. Research has shown that this leads to “productive failure,” i.e., errors in problem solution that can be used to provide feedback that increases conceptual understanding, fluency with specific procedures, and transfer to novel problems (Kapur 2014).

An example of this type of finding is reported by Kapur (2012) in an investigation of teaching high school students the concept of variance. Two groups were equated on ability by pre-testing, after which one group received direct instruction on established solution concepts and procedures, while the other group worked in small teams collaboratively to try out various representations and solutions (the error-producing group). Subsequently, the collaborative groups were provided direct instruction, and both groups then given novel problems to solve.

Analysis of problem-solving process measures derived from audio transcripts and worksheets showed that the error-producing group generated more problem solutions than the direct instruction group, and that transfer of learning to new problems was superior.

In order to address imbalances in group member participation in productive failure settings (i.e., errors are not discouraged), research by Westermann and Rummel (2012) investigated the utility of providing structured support for collaborative learning prior to direct instruction. This study was conducted with college students in the context of re-learning, integration, and application of various mathematical concepts to solve novel problems. Collaboration support was provided by a script that encouraged students to alternate roles in thinking, asking, and understanding (TAU)—in which one student would, for example, articulate a particular operation and the other student would ask for reasoning. Following the collaboration, a direct instruction session was provided. This was compared with a second group of students who received direct instruction only. Teachers who scored the math problems solved by both groups were blind to the experimental condition. The results indicated that as the TAU collaborative group became familiar with the process, their performance improved and became superior to that of the direct instruction group.

Error management training is defined as learning a task through exploration, with minimal guidance, and encouragement of errors as learning opportunities (Keith and Frese 2008). Meta-analysis of studies using error management training compared to procedurally-focused training in which errors were minimized suggest that transfer of learning to new tasks is superior with error management. Most of the studies in the Keith and Frese (2008) analysis involved learning various kinds of software application packages and used model solutions to novel application problems as an outcome measure. This limits the ability to generalize to specific work performance procedures.

However, work reported by Dyre et al. (2017) suggests that error management training for conducting ultrasound examinations of fetuses can enhance performance beyond simple procedure-oriented training. In this study, advanced medical students with no prior experience in sonography were instructed via a nine-module simulator course in the fundamentals of sonographic evaluation of fetuses, including weight estimation, head and abdominal circumference, and femur length. The error management training group was encouraged to explore different pathways within the simulator modules, not to restrict themselves to the specific topical order, and to view errors as opportunities to improve their skills. The procedure-oriented group was instructed to complete the module elements in the order presented and to try to minimize errors. Subsequent evaluation of the students on live-patient sonography evaluations showed that the error management training students scored significantly higher on a validated scale of structured ultrasound exam assessment and showed a non-significant trend toward more accurate measurements.

Another approach to error management training is the use of scenario reviews. Ellis and Davidi (2005) evaluated after-action reviews of failed and successful navigation exercises performed by military trainees. The dependent measure in this study was the quantity of concepts in mental maps drawn by the trainees to describe the events reviewed. Initially, failure scenarios provided more detailed mental maps than successful events. However, over time, the successful events were also analyzed in greater detail, suggesting that learning from failure can enhance trainee understanding of the elements of successful performance. Joung et al. (2006) observed a similar effect of reviewing failure scenarios of incident command in firefighting, i.e., ability to generate more numerous alternative actions, but did not see a corresponding change with success scenarios.

Feedback is typically considered a form of “summative” assessment, i.e., an end goal of the instructional process. However, the results of the studies of delayed instruction and error management training suggest an enriched interpretation—feedback as a constructive process, which can actively engage the learner in self-monitoring to improve learning. Instead of being a static process, feedback can become a continuous social process. Through the use of supportive materials such as learning criteria, checklists, and reflection sheets, learners can be engaged on a more continuous basis and develop preparatory strategies for subsequent training (Hattie et al. 2017). At the level of cognitive structures, the work on delayed instruction and error management training parallels results described in Appendix A, in which generation of responses prior to learning, elaboration and self-explanation of material, and correction of high confidence errors all lead to enhanced retention on tests of memory.

B.3.3.4 Relevance for NDE Training

There are numerous opportunities to provide feedback during NDE training, from standard classroom-based knowledge acquisition, to the less structured but more realistic supervised field experience. It is in this latter area that the principles discussed in this section are most applicable, particularly since NDE is a procedure-oriented job. Task-specific feedback would address the components of performance, e.g., was equipment assembled properly, weld profile accurately measured, etc. Process-oriented feedback might address such areas as *how* the weld profile was measured and combining individual scan steps to complete the entire inspection. These components would eventually combine during field experience to the point where the trainee would be able to self-assess their own performance with a backup check provided by the supervisor. This was reflected in interview comments from our earlier task analysis work (Sanquist et al. 2018), e.g., “You have to constantly watch yourself and be on guard not to complete the scan too early, and to ensure you have achieved full coverage.” Personal characteristic evaluations are also relevant here, in that there are team performance and cooperation issues that some examiners are better at than others; our earlier research suggests that this information is used in making assignments to work with others rather than trying to specifically train interpersonal skills. It is also apparent that feedback provided by PDI qualification testing is very non-specific, i.e., “pass/no-pass”—and is thus unlikely to facilitate skill development.

To the extent that work with practice samples and supervised field experience provide opportunities for exploration and learning from errors, the findings from the research on feedback and its elaborations should be incorporated. It is conceivable, for example, to engage a dyadic or collaborative process with multi-flaw practice specimens, in which a process akin to the Think-Ask-Understand script (described above) is used to promote exploration. Following this up with direct instruction from a supervisor would help to reinforce self-generated solutions and to correct errors in a non-critical environment.

B.3.4 Active Learning

B.3.4.1 Background

Active learning is a term that is used to express learner engagement in training, as distinct from being a passive recipient of information. The theoretical underpinnings of active learning derive from *constructivism*, a philosophy of learning based on the idea that learners construct their knowledge on the basis of experimentation and discovery, and generative theories of learning that focus on development of semantic associations and elaborations of knowledge structures. The reality of education and training is that instruction methods are often dictated by tradition and immediate circumstances, and that instructional techniques can be implemented in a variety of ways.

Studies of computer-oriented tasks analogous to workplace activities, such as operating word processor programs, display monitoring, and electronic information search, have suggested that active learning results in superior outcomes (Bell and Kozlowski 2009). These studies are conducted within the framework of exploratory learning that models workplace training, i.e., they use procedure-based tasks and do not rely on core discipline concepts. As such, they do not address the more important issues that underlie the development of technical expertise, i.e., learning basic content, developing problem-solving strategies, and integrating content into coherent knowledge structures. To address the role of active learning in these areas, it is necessary to draw from the field of instructional and educational psychology.

There is continuing debate among researchers regarding the extent to which minimally-guided instruction versus more structured active methods are most appropriate (Mayer 2004; Kirschner et al. 2006). There is general consensus, however, that the extent to which students actively engage with instructional material—whether through self or guided exploration, cooperative exercises, discovery, or other methods—is a key determinant of developing deep understanding of material and ability to transfer knowledge to new situations. An example is provided by research reported by Sawyer et al. (2017), which compared writing-to-learn exercises using text, cooperative learning with discussion, and multimedia with evaluative writing. Quiz results showed that the multimedia with evaluative writing yielded the highest scores, followed by cooperative learning, and text-based writing. Meta-analytic research comparing student performance and failure rates in science, technology, engineering and mathematics (STEM) classes using active vs. passive learning class activities showed that performance was improved by ½ standard deviation, and failure rates decreased from 34% to 22% (Freeman et al. 2014). This is also consistent with the findings from laboratory research on memory, i.e., more enriched cognitive processing—through elaboration, testing, and other methods—leads to better retention.

Active learning is a very broad term, subject to many different interpretations and implementations in learning contexts. As a consequence, instructors face a confusing array of techniques and ways to implement. This section discusses a recent theoretical formulation of active learning (Chi and Wylie 2014) as a way to present ideas about how students may be cognitively active through various levels of behavioral engagement. Specific approaches that have been demonstrated as superior to passive learning will be described that can provide a foundation for subsequent discussions of observational learning, simulation and multimedia, and tutoring.

B.3.4.2 ICAP: An Integrative Theory of Active Learning

Chi and Wylie (2014) have developed a framework of active learning that is meant to address the various definitional disparities in the field and to provide structure and concrete behavioral guidance to instructors for implementation. The framework is based on graduated levels of behavioral engagement, as defined below:

- *Passive* – learners are oriented toward receiving information without doing anything else overtly related to learning, e.g., simply watching or listening.
- *Active* – learners engage with material through some overt motor act, such as note taking or manipulation of an object.

- *Constructive* – learners engage with material beyond the information given, for example by generating self-explanations, concept mapping, or comparing and contrasting cases.
- *Interactive* – involves dialogues (between humans, or human and computer) that use constructive engagements and alternating between participants, as in asking questions, making explanations, elaborating on another person's contribution, criticizing by requesting justification, etc.

The learning outcomes for each of these modes of engagement are ordered, such that Interactive > Constructive > Active > Passive; hence, the acronym ICAP. This order reflects a general increase in knowledge, from passive where information is simply stored and can be recalled under very specific circumstances, to interactive, in which new knowledge structures are created and allow application of instructional material across a range of new situations.

In their discussion of this model, Chi and Wylie (2014) suggest that learning mode (ICAP) is orthogonal to a specific instructional task; for example, in observing a video (an instructional task), students may simply watch it and not do anything else (passive); they could stop, start, and rewind for playback (active); they could generate explanations of various concepts in the video (constructive); and they could debate with a peer about content and meaning (interactive). The boundaries between modes are flexible, and some instructional activities may be difficult to classify; further, the way individual students implement engagement activities may result in different modes. For example, in generating self-explanations, one student may integrate across concepts (constructive), while another may simply repeat instructional text more or less verbatim (active). Table B.2 compares the different learning modes in terms of knowledge-change processes, the expected changes in knowledge, the cognitive outcomes, and overall learning outcomes. It can be seen that as the learning mode moves from passive to increasingly active, knowledge elements interact and combine, create new knowledge, with the corresponding ability to not only recall information, but apply it in new contexts. The overall learning outcomes can be contrasted in terms of increases in levels of understanding and the ability to develop new ideas. The general progression of learning outcomes from passive to interactive parallels the stages of development of expertise discussed in Appendix A.

Table B.2. Example activities, knowledge-change processes, knowledge changes, cognitive outcomes and learning outcomes (based on Chi and Wylie 2014).

Category *Characteristic*	Passive *Receiving*	Active *Manipulating*	Constructive *Generating*	Interactive *Dialoguing*
Example Activities	Listening to explanation, watching video	Taking verbatim notes, highlighting sentences	Self-explaining; comparing and contrasting	Discussing with a peer, drawing a diagram
Knowledge change processes	Isolated "storing" processes in which information is episodically in encapsulated form without embedding it in a relevant schema, no integration	"Integrating" processes in which the selected and emphasized information activates prior knowledge and schema and new information can be assimilated in the activated schema	"Inferring" processes include: integrating information with existing knowledge; inferring new knowledge; connecting, comparing & contrasting different pieces of new information to infer new knowledge; analogizing, generalizing, reflecting on conditions of a procedure; explaining why something works.	"Co-inferring" processes involve both partners taking turns mutually creating. This mutuality further benefits from opportunities and processes to incorporate feedback, entertain new ideas, alternative perspectives, new directions, etc.
Expected changes in knowledge	New knowledge is stored, but stored in an encapsulated way	Existing schema is more complete, coherent, salient and strengthened	New inferences create new knowledge beyond what was encoded, thus existing schema may become more enriched; procedures may be elaborated with meaning, rationale and justifications; and mental models may be accommodated; and schema may be linked with other schema.	New knowledge and perspectives can emerge from co-creating knowledge that neither partners knew.
Expected cognitive outcomes	Recall: knowledge can be recalled verbatim in context (e.g., reuse the same procedure or explanation for identical problems or concepts)	Apply: knowledge can be applied to similar but non-identical contexts	Transfer: knowledge of procedures can be applied to a novel context or distant problem	Co-create: knowledge and perspectives can allow partners to invent new products, interpretations, procedures, and ideas.
Learning outcomes	Minimal understanding	Shallow understanding	Deep understanding, potential for transfer	Deepest understanding, potential to innovate novel ideas.

There is substantial empirical support for the ICAP model, both on the basis of pairwise contrasts between learning modes using relatively discrete tasks such as note taking versus none (Chi 2009), to more elaborate studies that test all modes of learning in a single study. An example of this latter type of research is provided by Menekse et al. (2013). This study involved undergraduate engineering students in a materials science course. The first experiment was a classroom-based assessment of different levels of learning activity upon quiz performance. For a unit on atomic bonding, for example, active processing entailed selection of the most likely material, bonding type, and processing method from a list of motorcycle parts; interactive processing entailed completing a partially drawn concept map through working with a partner; for a unit on crystal structures, constructive processing involved identifying the locations of atoms in a unit cell on a worksheet. Learning outcomes were assessed by performance on quiz items that required differing levels of cognitive processing, i.e., responses based on verbatim aspects of material covered in the course, responses based on integrating multiple concepts from course material, and responses based on drawing inferences about new problems related to course material. The general pattern of results showed that as the degree of learning activity increases, student performance improves. A follow-up laboratory study using similar materials with a greater degree of experimental control showed a specific ordering of learning impacts, with interactive learning yielding best performance, followed by constructive, active, and passive as shown in Figure B.5.

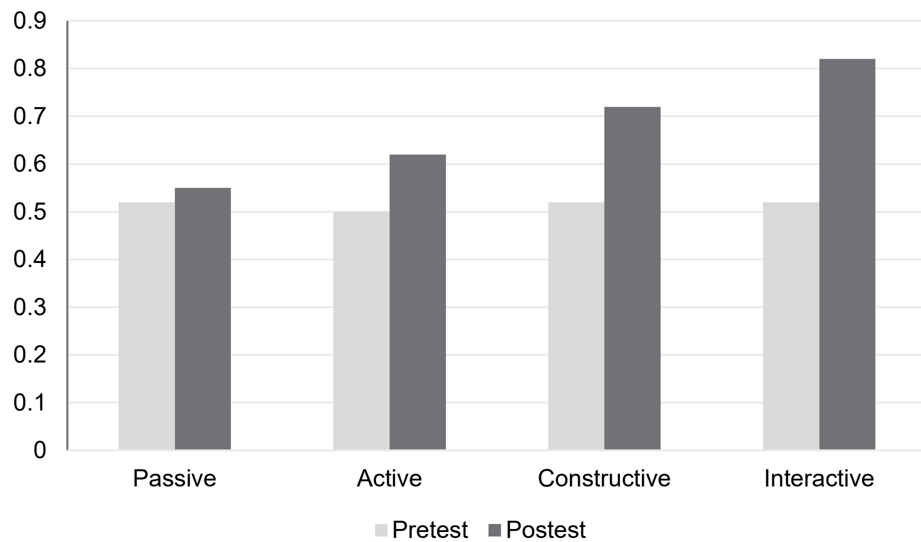

Figure B.5. Pretest and posttest performance by mode of learning engagement (based on Chi and Wylie 2014).

B.3.4.3 Instructional Techniques for Active Learning

This section discusses two techniques for active learning that have been shown to yield effective learning outcomes, which may be applicable to NDE training in appropriate circumstances. For each technique, we provide a brief description of the method, illustrative research findings, and guidelines for application.

B.3.4.3.1 Instruction based on examples

Students who study worked examples, such as algebra problems in math courses, show greater learning than students who simply work on solving equivalent problems. This effect has been demonstrated in numerous experiments across a variety of topical material (Renkl 2017).

An integrative theory of example-based learning has been developed by Renkl (2014) that includes four stages. In the first stage, learners acquire basic declarative knowledge about a domain—this is referred to as *principle encoding* and involves traditional instructional techniques. For NDE, this might entail learning about basic properties of materials and stresses. In the second stage—*relying on analogs*—learners begin to solve problems in the subject domain by studying examples of solved problems; e.g., a worked example of the stages of solving an algebra problem, or in the case of NDE, a flaw characterization. These examples are used as analogs to apply similar operations to new problems, rather than relying solely on application of domain knowledge principles. As skill is developed, learners begin to *form declarative rules*; they have attained a stage of verbalizable steps on how to approach problems. In the final stage, learners engage in *fine tuning, automation, and adaptation*; i.e., they learn to classify problems based on structure and apply rules accordingly.

The effectiveness of learning by example is thought to be based on the reduction of cognitive load afforded by examples. During the early stages of learning, when understanding of subject domain principles is shallow, studying examples helps learners to incorporate strategies that refer to the basic principles, and reduces reliance on superficial approaches more akin to trial-and-error. Further, studying examples facilitates the development of self-explanations, particularly when prompted, in order to reinforce subject domain principles (Renkl 2017).

Guidelines have been developed by Renkl (2017) for example-based instruction, synthesizing numerous experimental studies. These guidelines include using self-explanations, structuring example sets in specific ways to illustrate principles, using building blocks with explanations to help understand specific solution steps, using errors as examples, and eventual “fading” of problem-solving examples to conditions of more practice.

B.3.4.3.2 Instruction based on self-explanation

Self-explanation of reasoning during problem solving has been found to enhance the performance of students on subsequent similar problems (Gagné and Smith Jr. 1962; Rittle-Johnson and Loehr 2017). The term was first used by Chi et al. (1989) who found in studies of physics problem-solving that subjects who generated more explanations verbally during their solution process were more successful. A considerable amount of research has since been generated to understand this effect (reviewed by Rittle-Johnson and Loehr 2017).

Several decades of research indicate that prompting learners for self-explanations can improve comprehension of material and transfer of skill to new problem sets. Most of the research has focused on math and science material in laboratory settings; one study (Molesworth et al. 2011) found that self-explanation of altitude violations by pilots in flight simulators led to a reduction of safety problems in subsequent testing.

Self-explanation is thought to enhance learning and performance by means of promoting knowledge integration and focusing attention on structural features of to-be-learned material. Knowledge integration involves, for example, linking problem solution steps to previously learned conceptual material—various individual facts and observations are consolidated into coherent knowledge schema. Self-explanation can also guide learners to focus on potential solution methods, rather than individual problem content.

While self-explanation can enhance learning and performance, variations in how it is implemented can lead to greater or lesser benefits. For example, self-explanations that focus on domain conceptual material in a problem-solving environment have been found to decrease calculations performed, thus reducing problem-solving effectiveness. Further, self-explanations need to entail correct procedures or inferences, otherwise they will be counterproductive.

Rittle-Johnson and Loehr (2017) have developed four guidelines for use of self-explanations in instructional settings. First, they suggest that learning material that involves substantial conceptual content and core principles is more conducive to self-explanations than material that involves recall of specific facts. Thus, basic science, math, and engineering material is more appropriate for use of self-explanation than, for example, history or literature. Second, use of self-explanations should be supported through training learners on how to structure high-quality responses—this will reduce native strategies that may rely too heavily on prior knowledge or lack sufficient elaboration to reinforce learning. Third, learners should be prompted to explain correct information in worked examples—this has been shown to enhance transfer of training; self-explaining student-generated problem solutions that are incorrect can impair performance compared to no self-explanation. Finally, learners should be prompted to explain why incorrect information is incorrect—this is related to learning from errors, as described above.

B.3.4.3.3 Additional instruction approaches

The two instructional approaches described above—learning by worked examples and self-explanation—appear to be the most amenable to adaptation in the NDE training environment, which can take a variety of forms, including traditional classroom, self-guided study, and on-the-job training. There are other forms of instruction that have received considerable research in the educational psychology domain, including cooperative learning (in groups) and learning through various forms of inquiry (student-guided and problem-based), but these tend to be geared more toward traditional school curriculum disciplines, and evidence concerning effectiveness is variable. The research on these various approaches is reviewed in the relevant chapters of Mayer and Alexander (2017).

B.3.4.4 Relevance for NDE Training

Development of technical expertise such as NDE inspection would seem to provide ample opportunities for incorporating the elements of active learning, from initial basic material instruction, to OJT in the field. In the early phases of basic content learning, development of active exercises that address the operational relevance of materials science would be appropriate, as the examples of Menekse et al. (2013) showed for an engineering course. Simple exercises such as matching material types to lists of components, or integrative discussions of component failures in notable accidents, can be much more effective for knowledge retention and integration than repetition and recitation of abstract content. If computer-based training is used for fundamental content, students might be asked to select appropriate explanations from a list (described in Rittle-Johnson and Loehr 2017, pp. 353-354). Similarly, in a field experience setting, supervisors can structure the activities of trainees in such a way as to yield constructive or interactive learning experiences. For example, using the results of an inspection to characterize potential flaw indications, the supervisor might ask the trainee to explain the steps and to justify conclusions, rather than simply carry out the procedure, and leaving the interpretation to the more senior personnel. Due to the emergent aspects of operational settings, developing active-learning exercises "on the fly" will rely on the creativity and willingness of supervisors to tap the learning potential inherent in field experience. It is likely that guidance for supervisors can be developed that can facilitate this process. A form of this learning approach is already in use, e.g., in performance demonstration testing, candidates are required to provide detailed explanations of their calls.

B.3.5 Observational Learning

The use of role models and video demonstrations of complex tasks is considered a form of observational learning. Observational learning occurs all the time in everyday life—most particularly with children observing parents and other adults. In the realm of technical training, observational learning is based on structured presentation of model responses in the setting or task of interest. Two basic forms of observational learning are discussed in this section—behavior modeling training (BMT) and demonstration-based training.

BMT is based on social learning theory (Bandura 1977) and involves four component processes: attention, retention, reproduction, and motivation. BMT is typically manifested by means of trainees observing models—such as a video of a person performing the tasks and skills to be trained, or a live demonstration. Attention to the modeled behavior can be influenced by the distinctiveness of the skills to be demonstrated, how they are presented (e.g., easiest to most difficult), and the learners' perception of the model in relation to themselves (e.g., similarities or differences in age, sex, race). Retention of the modeled behaviors is related to establishing long-term memory, which has been shown to improve with mental practice (Taylor et al. 2005). Reproduction of the modeled behaviors serves as physical practice and permits feedback from trainers to further motivate the trainees. BMT has been applied to a range of skills, initially focusing on managerial interpersonal interactions, and extended to technical skills such as operating a computer system or software application.

Taylor et al. (2005) conducted a meta-analysis of 119 studies using BMT to train a wide range of skills. Outcome measures suggested that BMT effects vary widely across studies and appear to have larger effects on declarative and procedural knowledge measures, and relatively smaller effects on job performance. Increased hours of training were associated with improved performance. Assessments of retention and performance over time suggested that declarative knowledge decays, while skills and job behavior were sustained after training. Using learning points stated in terms of rules (if-then statements) during the initial presentation of BMT enhanced the development of procedural knowledge and skills.

BMT is a somewhat elaborate undertaking in terms of requirements—the knowledge and skills to be trained need to be analyzed into components appropriate for model scripting, then recordings made or demonstrations presented, followed by individualized reproduction and feedback. Further, there is substantial variation in how BMT is actually carried out—with some approaches using learning points, other approaches using retention aids, and yet others using task performance as a basis for assessment without prior practice and feedback. A more flexible conceptualization of observational learning is provided by the framework of demonstration-based training (DBT), discussed in a recent review by Rosen et al. (2010).

DBT is similar to BMT in that observing a demonstration of the target behaviors and skills is a core element. A demonstration is "a dynamic example of partial- or whole-task performance or of the characteristics of a task environment that illustrates with video, modeling or any visualization, the enactment of targeted knowledge, skills and abilities" (Rosen et al. 2010, p. 597). The context surrounding the demonstration comprises the variable aspect of DBT—i.e., the extent to which information is presented along with the model example, or whether various activities take place before, during, or after the demonstration.

Information that is presented along with the demonstration is considered as passive guidance or support. This may consist of instructional narratives or behavioral summaries that the trainee reads or listens to in lecture form. The intent is to focus attention during the demonstration—a form of priming. Instructional narratives can be provided along with the demonstration and can be used to describe *covert* aspects of performance, i.e., the model's thought processes, or the reasons for certain overt behaviors.

Rosen et al. (2010) also describe *activities* that can be undertaken in the context of a demonstration. This can involve preparation to view, such as prompting on specific elements to watch for, or concurrent tasks such as taking notes or coding behaviors performed by the model. Further, trainees can engage in analysis of activities observed after the demonstration (retrospective) or consider application of what they have seen to future tasks (prospective). Examples of retrospective activities include mental rehearsal of the skill demonstrated, self-generated learning points (similar to self-explanation), and instructor-guided discussion. Examples of prospective activities include goal setting for use of the demonstration content (e.g., where and when), generation of rule codes (conditions) for different task contexts, and self-generated practice activities. A variety of research evidence is reviewed to suggest the effectiveness of these various techniques, which share many common elements with learning by example and self-explanation.

B.3.5.1 Relevance for NDE Training

The practical aspects of NDE training rely heavily on observational elements. The field experience requirements, for example, present an extended opportunity for observational learning. The extent to which elements of BMT and DBT are used in field experience are unclear, but it is likely that they are applied at least informally (as in the medical training adage "see one, do one, teach one" in reference to surgical procedures). Research evidence indicates that BMT and DBT can be enhanced by various concomitant activities, including use of learning points prior to observing the model behavior, mental practice following observation, and anticipatory activities such as goal setting and rule formation.

BMT and DBT should be thought of as flexible frameworks within which observational learning may occur during field experience. While it may not be feasible to develop fully scripted and recorded behavioral models, supervisor actions such as component skill demonstration followed by trainee reproduction and feedback can capture the essential elements of active learning implemented by means of observation.

B.3.6 Learning by Visualization and Multimedia

B.3.6.1 Background

Traditional instruction has been largely lecture- and text-based. Textbooks have incorporated pictures as complementary or essential material since Comenius introduced *Orbus Pictus (The World in Pictures)* in 1658 (described in Mayer 2017). With the advent of visual media such as television, motion pictures, and computers, it is now possible to create and present high-quality visual representations of most types of learning material. Visualizations are visual-spatial representations that include pictures, drawings, videos, and animations, and together with accompanying text comprise a multimedia instruction message that is intended to promote learning (Mayer 2017). This section discusses research findings that address effective ways to design and present visualization and multimedia.

The multimedia instruction hypothesis suggests that people learn better from words and pictures than from words alone. This proposition has been confirmed in numerous experiments. In a relatively simple study comparing learner performance on tests of knowledge concerning vehicle braking systems, Mayer (1989) found that performance was better for text accompanied by illustrations, than by text alone. Additionally, labels that linked aspects of the illustration to concepts in the text were found to be important for directing attention to explanative text and promoting transfer to making inferences about system malfunctions. Similarly, Kalyuga et al.

(1998) found that integrated text and illustrations increased knowledge and fault-finding performance of junior electrician apprentices in tasks related to circuitry.

Multimedia learning is based on the structure of the human cognitive system (discussed in Appendix A), which represents the flow of information through various stages of processing, as shown in Figure B.6. Words constitute verbal information that is processed through separate organs and mental representations (channels) than pictures.

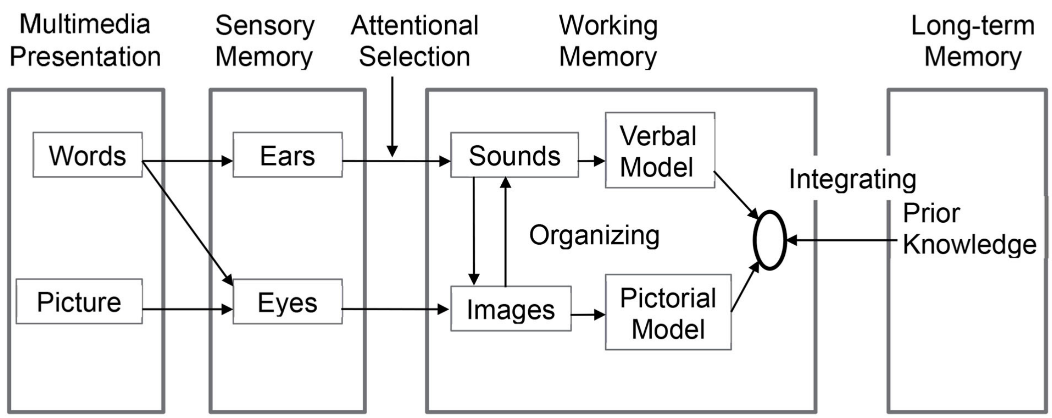

Figure B.6. A cognitive model of multimedia learning based on (Mayer et al. 2001).

The information flow depicted in Figure B.6 portrays a *selection* of words and images from the input stream, *organization* of that information into mental models, and *integration* of the information forms links with prior knowledge in long-term memory. Selection involves paying attention to relevant information and reducing extraneous processing. Organizing builds coherent representations and entails processing of information that is essential for learning. Integrating builds connection to what is already known, and involves generative processing, i.e., generating connections to existing knowledge. Effective multimedia instruction facilitates selection by reducing extraneous and enhancing relevant information, limiting essential processing to what is most important, and fostering generative processing by making connections between verbal and visual material.

A considerable number of variables influence learning by multimedia; the following discussion addresses the primary research findings and resulting design principles for multimedia instruction. A convenient way to encapsulate the 30 years of research findings addressing multimedia instruction material is in terms of design principles that influence the various types of cognitive processing discussed above. The research findings supporting the following design principles are discussed in Mayer (2008, 2017).

B.3.6.2 Selection – Reducing Extraneous Processing

Focusing selective attention with multimedia functions can reduce extraneous cognitive processing. This can be achieved in five ways:

1. *Coherence* – reducing extraneous material. It is tempting to add “flavor” or seductive details, but this reduces focus on the most important material.
2. *Signaling* – when certain information cannot be totally eliminated, the important material can be highlighted, i.e., signaled by adding overviews or headings.

3. *Redundancy* – visual and narrative material presented orally result in better learning than visual, narrative, and text overlays.
4. *Spatial contiguity* – presenting words near the appropriate pictures results in better learning.
5. *Temporal contiguity* – simultaneous presentation of words and appropriate images results in better learning.

B.3.6.3 Organization – Enhancing Essential Processing

Essential processing can place excessive demands upon the learner despite efforts to reduce extraneous material. The following design principles provide ways to manage information load to enhance essential processing:

1. *Segmenting* – sheer volume of material can overwhelm learners; by breaking continuous lessons into smaller segments, learning is enhanced.
2. *Pretraining* – making learners familiar with key terms and elements prior to presentation of instructional material helps to focus attention.
3. *Modality* – learners perform better when information is presented in graphics and spoken text, instead of visually presenting the text on screen.

B.3.6.4 Integration – Fostering Generative Processing

Knowledge connections between modes of learning material presentation and with prior knowledge are established with generative processing. The following design principles provide ways to foster this:

1. *Multimedia* – connections between visual and verbal material are established when words and pictures are used in combination.
2. *Personalization* – using conversational style (“you” instead of “the”) results in better learning transfer.

B.3.6.5 Applications

Most of the research upon which the aforementioned design principles are based has been conducted in a laboratory setting. Two studies of medical education conducted in a classroom setting provide evidence that the multimedia design principles can be applied to medical course content and yield superior learning as measured in post-instruction tests (Issa et al. 2011; Issa et al. 2013).

Learning material for the classroom study was derived from a 1-hour lecture on shock syndrome, in a required series of lectures on various topics delivered to third-year medical students. The original learning material had been developed as PowerPoint slides prior to the research study and used in previous lectures. These slides were revised by the lecturer and an educational specialist using the Mayer multimedia principles. The revised slides were reviewed by another content expert to ensure that all necessary material was still present in the revised slides.

Students took pre- and post-lecture tests to assess prior state of knowledge, subsequent retention, and transfer. In follow-up research, students also took tests of delayed retention and transfer at 1 and 4 weeks after the lecture. The results showed that the slides modified by multimedia design principles yielded superior retention immediately after the lecture and on the subsequent delayed tests, with similar results for questions involving knowledge transfer.

B.3.6.6 Relevance for NDE Training

The nature of NDE is multi-sensory, i.e., it involves examiners using vision, physical sensation and movement, and verbal communication to accomplish the job. Further, the nature of the fundamental material that needs to be mastered for qualification is dynamic and structurally complex, thus lending itself to multimedia representation. The material discussed in this section suggests that specific design principles can be employed to enhance the effectiveness of multimedia instruction; conversely, the relatively easy ability to present an overabundance of information can tempt instructors and multimedia designers to insert more material than necessary to convey the to-be-learned material (death-by-PowerPoint). Thus, adhering to design principles for multimedia material can help to focus, organize, and integrate complex material.

Research indicates that there are tradeoffs between levels of expertise and type of information presentation. Multimedia tends to enhance learning for relative novices but is less effective for more experienced individuals. Application of the multimedia design principles to NDE training would best be initiated by a selective review of fundamental material used by trainees to prepare for initial qualification exams. Based on results of such a review, targeted application of multimedia implementations could be developed and assessed.

B.3.7 Simulation

B.3.7.1 Background

A level of active learning that extends beyond static or animated visualization is that of simulation. Simulation entails a computational model of a system or a process—such as aircraft flight, patient anatomy, or a physiological process such as blood flow—that is embedded in representative physical models with which students interact. The training advantages of simulation are numerous, and include potentially lower cost, the ability to create scenarios and conditions that would be dangerous in the real world, the ability to repeat conditions or tailor training to specific individual needs, and the ability to provide and receive feedback. Learning by simulation embodies several psychological principles that are manifested in field experience, including:

- Knowledge of results—trainees see the effects of their actions upon simulated processes represented in displays
- Perceptual learning—trainees have the ability to extract information from stimulus patterns, e.g., dynamic stimuli such as out-the-window views
- Stimulus-response learning—trainees make connections between, for example, a pattern display and the type of action they must take (Adams 1979).

Simulator fidelity varies considerably, but it is generally agreed that fidelity is less important than the details of the training program employed (Caro 1973). While simulation has been employed historically for training practical skills such as flying or surgical procedures, more recently it has been employed in educational settings to teach scientific concepts. The fundamental question involved in simulation concerns transfer; i.e., the extent to which simulator training improves performance in actual job circumstances. This can be assessed in various ways, including error reduction and time savings for training in real-world performance. Answers to these questions have a bearing on the extent to which simulator training can be used to substitute for field experience.

Because simulator substitution percentages are sometimes codified in regulations for licensing—as in aviation—it is important to have a quantitative definition for transfer of training from simulation to operations. This helps to establish an understanding of exactly how much simulator time is expected to improve operational knowledge and skills, and thereby reduce training time needed in the field. The basic transfer of training computation is provided and illustrated with hypothetical data by Roscoe (1971). Table B.3 shows notional data for two groups of students—Group 1 engaged in 100 hours of prior study (e.g., simulation) and required 100 hours of additional training to pass a criterion test. Group 2 engaged in no prior study and required 200 hours of training to pass the criterion test.

Table B.3. Hypothetical transfer of training data; adapted from Roscoe (1971).

Group	Hours of Simulator Study	Hours of Operational Training Required to Pass Criterion Test
1 – Experimental	500	100
2 – Experimental	250	100
3 – Control	0	200

The basic transfer computation, referred to as "syllabus reduction," is as follows:

$$\frac{(Y_c - Y_e)}{Y_c}$$

where:

Y_c = time, trials, or errors for the control group
Y_e = time, trials, or errors for the experimental group

Substituting the values in Table B.3 yields the following computation:

$$\frac{200 - 100}{200} = \frac{1}{2} = 50\%.$$

The results are the same for experimental group 2. It can be seen from the table and calculation results that the syllabus reduction ratio does not account for hours of simulator study, and that there is a considerable difference between the two groups.

An alternative metric is the transfer effectiveness ratio (TER), which quantifies transfer in terms of the amount of training time saved to the amount of prior training time (e.g., simulator). In the case of the sample data above,

$$\text{TER} = \frac{(Y_c - Y_{e1})}{Y_{s1}} = \frac{200 - 100}{500} = 0.2$$

where Y_{s1} is the amount of simulator time required for the first experimental group. In this case, 1 hour of simulator time reduces operational training time by 20%.

Making the same calculation for experimental group 2 yields

$$\text{TER} = \frac{(Y_c - Y_{e2})}{Y_{s2}} = \frac{250 - 100}{250} = 0.6.$$

Thus, reducing operational training time by 60%.

Ratios greater than 1.0 result when the amount of criterion (operational) training time saved by prior simulator training is equal to or greater than the amount of time spent in prior training. In terms of simulator usage, a TER > 1.0 results when time saved in operational training is greater than the amount of time spent in the simulator. A TER < 1.0 occurs when training time saved is less than simulator time invested.

Thus, in the example, 500 hours of prior study reduces the hours of operational training required to pass a criterion test by 50%, and increases the training time saved by 20%, i.e., an hour of simulator time saves 12 minutes of operational training time. In experimental group 2, an hour of training time saves 36 minutes of operational time. These types of computation have been used in many studies of transfer in the aviation community, primarily in the 1960s–mid-1980s and are discussed below.

Roscoe (1971) interprets data from the psychology of learning showing negatively accelerated learning functions to suggest that simulator transfer is likely to be incrementally less effective over time. That is, the first hour of simulator training will have a larger effect upon operational time saved, the second hour will be less effective, and so on. A graphical representation of the relationship between percent transfer (syllabus reduction) and the TER for notional flight simulator training data is shown in Figure B.7. The curves in Figure B.7 show hypothesized relationships between percent transfer and incremental transfer effectiveness, based on data extracted from earlier simulator studies. It can be seen that the effectiveness of simulator training, in terms of percent transfer, is steep during the first several hours then begins to level off. The cumulative transfer effectiveness function begins to level off in the later hours of ground simulator training, the subsequent benefit of further simulator hours is considerably diminished, and the incremental change in flight hours saved is reduced to near zero. In the sample data shown, 1 hour in a ground trainer saves 1.4 hours in actual flight training time (14% transfer, 1.4 transfer effectiveness); 5 hours of simulator time saves 5 hours of actual flight time (50% transfer, 1.0 transfer effectiveness); and 15 hours of simulator time saves 7.5 hours of actual flight time (75% transfer, 0.5 transfer effectiveness).

	GROUND TRAINER HOURS, X															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Flight Hours, Y	10.00	8.60	7.44	6.48	5.68	5.01	4.45	3.98	3.59	3.27	3.01	2.81	2.66	2.56	2.51	2.50
Flight Hours Saved, $Y_o - Y_x$		1.40	2.56	3.52	4.32	4.99	5.55	6.02	6.41	6.73	6.99	7.19	7.34	7.44	7.49	7.50
Percent Transfer, $\frac{Y_o - Y_x}{Y_o} \cdot 100$		14.0	25.6	35.2	43.2	49.9	55.5	60.2	64.1	67.3	69.9	71.9	73.4	74.4	74.9	75.0
Incremental TER, $\frac{Y_{x-\Delta x} - Y_x}{\Delta X}$		1.40	1.16	0.96	0.80	0.67	0.56	0.47	0.39	0.32	0.26	0.20	0.15	0.10	0.05	0.01
Cumulative TER, $\frac{Y_o - Y_x}{X}$		1.40	1.28	1.17	1.08	1.00	0.93	0.86	0.80	0.75	0.70	0.65	0.61	0.57	0.53	0.50

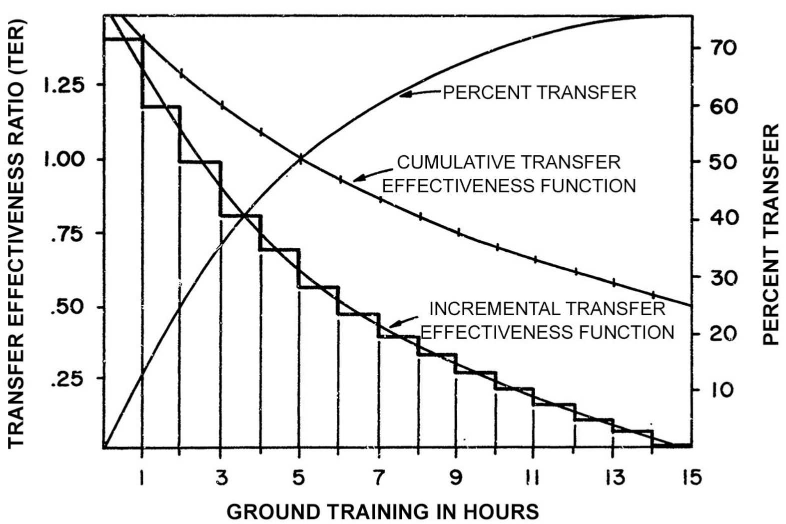

Figure B.7. Relationship among transfer measures based on hypothetical data for aviation ground trainers (Roscoe 1971).

The quantitative relationships between simulator training and impact upon operational training time saved implied by Roscoe's (1971) analysis have not been investigated in detail due to the difficulty of implementing the factorial combinations required for measurement and validation. However, the studies reviewed below suggest that initial training is substantially enhanced by prior simulator practice, suggesting that the early learning impact suggested by the transfer function analysis is accurate. More generally, the concepts suggested by Roscoe's (1971) analysis provide a basis for evaluating proposed simulator substitution regimens and their implied impact.

B.3.7.2 Aviation

Studies of transfer of training in aircraft simulators have been conducted for many years in an attempt to evaluate issues associated with fidelity requirements, specific task training, and the extent to which such training can be used to reduce the expensive use of actual aircraft flight time for training. An extensive compilation of transfer studies was reported by Diehl and Ryan (1977), in which research addressing simulator training for general, commercial, and military aviation were assessed. The studies included designs evaluating the impact of simulator training upon performance in operational aircraft. Twenty-five studies were evaluated, and twenty-four reported positive syllabus reduction results, ranging from 1%–93%. Despite the wide range, the data clearly indicate positive transfer of training from simulators to operational aircraft, most with a substantial reduction in the number of aircraft hours required to meet proficiency standards. A subsequent review of military aviation transfer studies between 1986 and 1997 (Carretta and Dunlap 1998) also found positive transfer from simulator to aircraft and suggested that simulators were effective for training landing, skill, bombing accuracy, and instrument flight control. Further, as the number of simulated sorties increased, transfer effectiveness decreased, leveling off at 25 missions.

The studies covered in the reviews described above were based on student and professional pilots who had already achieved some degree of operational aircraft experience. Several more recent studies have assessed ab initio (from the beginning) training in simulators for inexperienced students and the extent to which there is transfer to operational aircraft. In a series of studies conducted for U.S. Army initial helicopter pilot trainees Stewart et al. (2002) reported positive transfer of simulator training for a variety of maneuvers. In one experiment it was shown that approximately 2 hours of simulator time could be used to eliminate 1 hour of operational aircraft time. In a substitution experiment, it was determined that 9 hours of simulator time could be substituted for 7.8 hours of aircraft time with no detrimental effect upon trainee proficiency measures in operational testing. In research conducted with students training for a private pilot license, Macchiarella et al. (2006) reported that a curriculum comprising 60% flight training device and 40% operational aircraft experience showed positive transfer of training on 33 of 34 maneuvers compared to a control group that received 100% of its training in operational aircraft. In a similar experiment, McLean et al. (2016) evaluated a combined simulator (20%–25%) and operational aircraft (75%–80%) curriculum on student performance at various stages of learning, including first solo, general flight progress test, and private pilot license proficiency. They found that the simulator group required more total training time (104 hours) versus (98 hours) to reach private pilot license proficiency, but fewer total hours in the operational aircraft (90 hours versus 95 hours). For each of the earlier stages of training, operational aircraft time was also reduced.

The use of simulation in training is addressed by a variety of Federal Aviation Administration (FAA) regulations, which specify operational flight hours and permissible simulator hours for different types of pilot licenses. For example, CFR Title 14 Chapter I Subchapter D Part 61 (*Certification: Pilots, Flight Instructors and Ground Instructors*) specifies that to obtain an air transport pilot certificate, the candidate must have 1500 hours of total time as a pilot. This includes 500 hours of cross-country flight time, 100 hours of night flight time, and 50 hours of time in the class of airplane for which the rating is sought. For the airplane type rating, a maximum of 25 hours of training in a full flight simulator is allowed if the training was conducted as a part of an approved training program. For the overarching requirement of 1500 hours, no more than 100 hours may be obtained in a full flight simulator (a substitution of less than 10%). Further requirements address instrument ratings, which permit a somewhat higher level of simulator substitution, wherein a candidate must log 50 hours of cross-country flight time as a pilot, 40 of which may be performed in a full flight simulator. Currently, for pilot training schools, simulator substitution is permitted to a maximum of 30% for flight time requirements for full flight simulators and 20% for "flight training devices" (lower fidelity) for commercial pilot licenses (14 CFR Appendix D to Part 141, *Commercial Pilot Certification Course*).

The regulation of pilot certification is complex and involves requirements that vary according to how and when aeronautical experience is gained, which also affects allowed simulator substitution percentages. In addition to the governing 1500 hours of flying time, type ratings for different kinds of aircraft must be obtained, most of which can be done in a simulator if the candidate already has jet aircraft flying experience. Beyond basic and type rating flying hours, 14 CFR Part 121 (*Air Carrier Certification*) contains additional requirements that must be achieved by pilots who have already been certified for air transport operations and aircraft type. This includes an additional 25 hours of flight time with a line check airman from the employer airline, and a check flight with an FAA inspector. Of particular interest for the issue of recurrent practice is the requirement that

> No certificate holder may use any person nor may any person serve as a required pilot flight crewmember, unless within the preceding 90 days, that person has made at least three takeoffs and landings in the type airplane in which that person is to serve. The takeoffs and landings required by this paragraph may be performed in a visual simulator approved under §121.407 to include takeoff and landing maneuvers. In addition, any person who fails to make the three required takeoffs and landings within any consecutive 90-day period must reestablish recency of experience as provided in paragraph (b) of this section. (14 CFR Part 121.439 (a)).

B.3.7.3 Medical Skills

Learning and practicing medical skills has a long history of using simulations of live patients, including practice on cadavers and live animal specimens for various surgical techniques, as well as "standard patients" engaged in role playing and mannequin-based approaches. With changes in medical education curriculum to include more emphasis on clinical skills training and less of an apprenticeship model, simulation has been used to fill gaps in development (Willis and Van Sickle 2015). Various review boards and guidance committees require the use of simulation training in surgical residency programs, and board certification requires passing tests that include skills demonstration on laparoscopic and endoscopic simulators. A considerable range of low-fidelity simulation tools are used in "boot camps" that prepare senior medical students for surgical residency and include techniques such as tying and suturing, airway management, central venous catheterization, chest tubes, and laparoscopy skills. Simulation and skill laboratories are now mandated by the Residency Review Committee for Surgery (Willis and Van Sickle 2015). Laparoscopic surgery has received the most attention recently in the literature in terms of transfer of training and effectiveness studies, several of which are reviewed below. Additionally, the area of minimally invasive surgery has a number of parallels to NDE examination, including device manipulation and imaging to determine necessary interventions.

Early laparoscopic "box trainers" consisted of an opaque box that approximated the size of the human abdominal cavity, and were equipped with prefabricated slits for inserting instruments and a camera on a flexible arm that shows an image on a video monitor.

A variety of training drills have been developed for use on box trainers, including manipulating small objects such as beans and suturing practice using standard sutures, needles, and sometimes tissue specimens to improve realism. The sensory feedback from these devices is similar to that obtained during surgery (Roberts et al. 2006). Costs for box trainers are relatively low, and their use has been shown to improve skills on surgical component tasks and transfer to live surgeries (Moulton et al. 2006; Dawe et al. 2014).

Virtual reality (VR) simulators use computer software to represent the 3D visual location of physical instruments that are manipulated during training. Software permits the simulation of a larger variety of tasks and tissue manipulations, including cutting, dissection, and diathermy (electrocautery). VR simulators also provide the capability to make recordings of procedures that include instrument locations, thus providing a feedback capacity (Roberts et al. 2006).

One of the earliest studies of VR surgical transfer of training was conducted by Seymour et al. (2002). This investigation compared groups of surgical residents who had received standard training for laparoscopy and VR training. The residents subsequently performed supervised laparoscopic removal of a gallbladder and were assessed by physicians blind to the type of training received. Recordings of the procedures were made and scored according to previously defined operative errors. Significantly fewer errors were made by the VR-trained group (average = 1 error per procedure for the VR group, 7.25 errors per procedure for the standardly trained group). Subsequent reviews of simulator training have confirmed positive transfer to the operative setting (Dawe et al. 2014). Further, individualizing the simulator tasks practiced by residents based on initial surgical assessments—referred to as "deliberate practice"—suggests that tailoring study regimens may be an effective way to build expertise and lower the cost of "one size fits all" simulator training (Palter and Grantcharov 2014).

A medical task that is similar to NDE inspection is the use of ultrasound for diagnostic imaging. Ultrasound is widely applied in specialties ranging from OB/GYN to cardiology and vascular assessments. The use of simulation for training in medical ultrasound is relatively limited compared to surgical procedures, based on issues associated with 3D reconstruction of complex anatomical structures, variations in tissue density, and the need to embed reflectors in an appropriate physical representation. Phantoms mimicking the relevant physical properties of real tissue tend to be the primary basis for current ultrasound training. Nonetheless, there are a variety of limited-scope simulators available for training (Blum et al. 2013). Thus, despite limitations in image fidelity, haptic feedback, and physical landmark representation, ultrasound simulators are beginning to be employed in medical training programs. Tolsgaard et al. (2015), for example, reported a study of physician residents who were trained conventionally (i.e., by supervised ultrasound performance on live patients) compared to a group trained clinically whose learning was examined using an ultrasound simulator. Follow-up assessment of both groups indicated that the physicians trained with the ultrasound simulator performed better on all measures of an objective examination assessment and maintained their skills for several months beyond the training period. Blum et al. (2013) suggest that enhancements in visualization techniques may help alleviate issues of image fidelity, and a balanced approach of procedural training with clinical application can be cost-effective.

Substitution studies have not been formally conducted for surgical simulation, but research reported by Gallagher et al. (2013) suggests that a TER of 7%–42% can be considered illustrative of the improvement in skill; that is, simulator training may reduce operational training by as much as 42% for novices and also seems to reduce error rates for experienced surgeons (7%). In nursing, some states permit substitution of simulation for clinical training hours, but no standard exists for a substitution ratio (Breymier et al. 2015).

B.3.7.4 Relevance for NDE Training

Proposals to reduce the amount of field experience to qualify as a Level II examiner and to permit substitution of laboratory exercises for most of the reduced hours is essentially an endorsement of simulation in some form. Laboratory exercises may consist of scanning on real samples with service-related or manufactured flaws or potentially a simulator that uses computer-generated or file-based imagery. The goal of the industry proposal is the same as that in other fields using simulation—to reduce extended duration operational training that is time-consuming and expensive. The domain with the most experience in this regard is aviation. While it is clear that training transfer can be used to reduce the operational requirements for proficiency, regulation permits a maximum substitution of 100 hours out of 1500 hours for air transport pilot certification (7%) and somewhat higher for added certificates, such as instrument qualification. Somewhat higher percentages are allowed for students within approved training programs, but not higher than 30%.

Medical practice is engaging simulation at an increasing rate, but the gold standard is still clinical experience. Given these relatively conservative approaches in domains with considerably more experience in training hours and potential surrogates, it is warranted to proceed with caution in NDE. The proposed modifications to ASME Code Section XI, Appendix VII amount to a potential reduction of field experience of up to 90%, far in excess of any substitution ratio now used in other domains. In terms of the TER, 320 hours of lab time would reduce field training time to 80 hours, a reduction of 720 hours. This represents a TER of 2.25; substituting 1 hour of lab/simulator time can reduce field training time by 2.25 hours, well beyond what has been reported in the literature or permitted for initial pilot certification. Further, while the difficulties of simulating clinical ultrasound may not be completely applicable in NDE because physiological parameters are not involved, a significant degree of fidelity testing would be warranted before incorporating simulation as a primary tool in NDE training.

B.3.8 Experience, Practice, and Deliberate Practice and the Development of Expert Skill

B.3.8.1 Background

In Appendix A, we discussed the stages of skill acquisition as proceeding from knowledge-based representations to rule-based applications, and finally to skill-based performance in which correct problem solutions are directly applied without a more deliberate or conscious process of assessment and selection. While this general framework is still applicable in many situations, more recent developments in the study of expert performance suggest a critical role of *deliberate practice*, as distinct from the simple accrual of practice and experience in a task. A further parallel from the studies of learning discussed in Appendix A involves the effects of time on retention. In general, the more time spent studying material, the better it is retained. However, many studies have shown that how the extra time is used makes a critical difference in retention—for example, self-testing instead of simply re-reading enhances retention. Similarly, for learners in whom study time is equated, those who distribute their study time retain more than those who study all at once.

The material discussed in this section is related to these fundamental findings in human learning, in that the distinction is made between simple practice, i.e., repetition of specific aspects of performance, and *deliberate practice*, i.e., repetition and drilling of increasingly more difficult aspects of performance in association with feedback. Research has accrued that suggests deliberate practice is what distinguishes individuals who show superior performance from those who are competent or proficient.

B.3.8.2 Practice and Experience in Simple Tasks

The earliest studies of the effects of practice on skill improvement were conducted on manual telegraphy operators (Bryan and Harter 1899). The findings indicated that speed of performance improved over a period of many weeks, progressing from stages in which individual letters were perceived to a point where more integrated units of words and parts of sentences could be anticipated. Performance appeared to increase for a period of time, followed by what was termed a “plateau,” i.e., a period of no improvement. Plateaus were initially ascribed to a learning process in which skills were becoming more organized and automated and were followed by subsequent improvement. Many years later, large-scale studies of code operators showed that improvement was continuous and that plateaus were an artifact of the original study circumstances (Fitts and Posner 1967).

Many studies have now been conducted evaluating improvement in speeded performance and error reduction in simple tasks such as 10-key data entry, tracing mirror images, and visual target scanning (Fitts and Posner 1967; Newell and Rosenbloom 1981). The most common finding is that performance continues to improve over a long period of time and that the relationship is best described as a power function, i.e., a log linear relationship. A practical illustration of the effect is shown in the work of Crossman (1959), who studied improvement in cigar-making speed (Figure B.6). In this study, speed of work continued to improve over a two-year period until it approached the limits imposed by the machine.

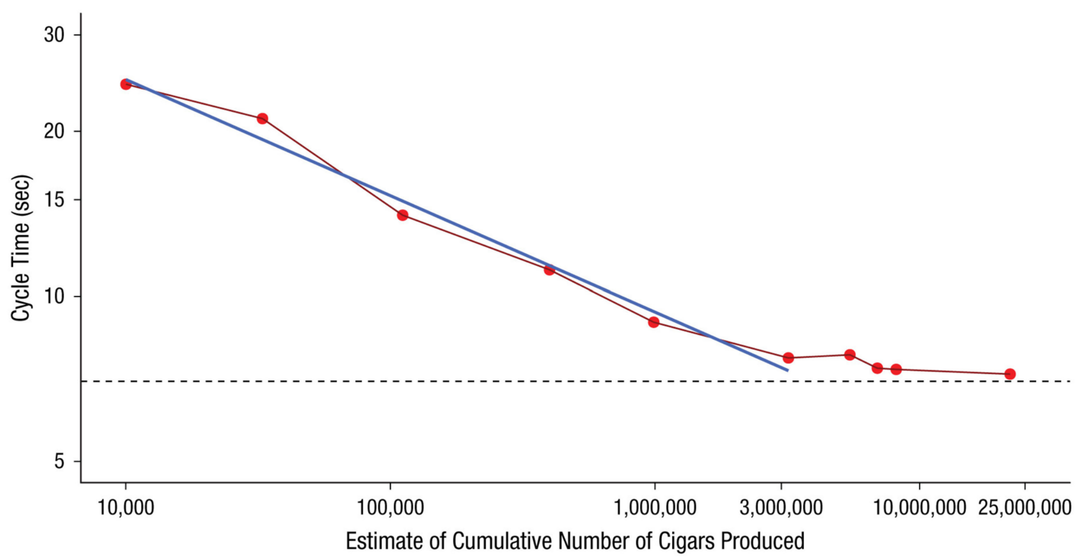

Figure B.8. Improvement in cycle time of cigar makers over time (Crossman 1959; from Gray 2017).

The relevance of simple task practice effect studies to the concerns of NDE training is that evidence suggests continued practice results in continuing improvement. The early ideas about “plateaus” of learning in which skill becomes more organized should be considered distinct from “asymptotes,” i.e., a flattening of the performance curve due to inherent physical limits. Current theory in psychology considers a plateau to be a limit which can be surpassed, rather than imposed by fundamental mechanisms (Gray 2017).

B.3.8.3 Practice and Experience in Complex and Professional Tasks

The effects of practice and job experience on more complex tasks such as driving a car, typing, language translation, auditing, and other domains are initially similar to the results obtained with simple tasks. Since the primary goal is to reach a level of acceptable proficiency, the initial activities usually involve avoiding errors and consciously executing required patterns. As performance improves, actions are performed with less conscious effort (automatically) and tend to stabilize at a plateau where no further improvement is observed. This is quite typical for recreational activities such as golf or tennis, where performance for an individual is stable even after many decades of playing (Ericsson 2018). Additionally, in various professions such as medicine, nursing, psychotherapy, and teaching, there is little relationship between years of professional experience and outcomes such as diagnoses, patient health, and student improvement. There is, however, aggregated evidence to indicate that specialization in various aspects of medicine or higher volumes of patient treatments is associated with better outcomes. The mere accrual of experience beyond the level necessary to become proficient does not necessarily lead to further improvements.

Despite the general tendency toward performance improvement plateaus, there are individuals who are considered to be superior performers based on various criteria such as peer or teacher nomination, international reputations, competitive status, and exceptional performance. The study of experts and comparisons to novices or lower tier performers was initiated in the 1940s by de Groot in studies of chess masters (discussed in Ericsson and Lehmann 1996). In a series of studies comparing different classes of players (e.g., novice, accomplished, and world-class), it was found that expert players generally were able to select the best moves during their initial perceptions of board layouts, performed more extensive analyses of potential future moves (as determined by verbal protocol analysis), and were able to better recall board positions when pieces were arranged in meaningful ways. Their memory for board positions, however, was no better than average players if the pieces were arranged in a random fashion. Developmental analyses suggested that chess masters start early in life, require at least 10 years to reach international competition status, and spend up to 4 hours per day analyzing published games of master players (Ericsson et al. 1993). Similar findings have been obtained for concert-level musicians, athletes, and expert typists. The distinguishing characteristic of the developmental experience of these experts is the amount of time spent in solitary, focused practice that yielded immediate feedback and additional opportunities for improvement.

A key aspect of studying experts is performance analysis in laboratory situations, sometimes using “think-aloud” protocols in which concurrent verbalization is performed as problems are solved using various memory and reaction time tasks. In general, it is found that experts exhibit a more detailed and extensive knowledge base than non-experts, are better able to recall domain-relevant material (e.g., case information, chess board positions), and respond more quickly on perceptual-motor tasks (Ericsson and Lehmann 1996). This general pattern has led theorists to suggest that experts, by practicing their skill, eventually adapt at higher levels to performance constraints in the component aspects of their work.

The work of chess masters, musical artists, and competitive athletes is aimed at peak performance during events such as concerts and competitions, and these events are supported by relatively continuous training activity that increases as the performance event approaches. Most jobs, including NDE, do not entail this type of solo performance, but instead involve expectations of acceptably proficient performance or better on a continuing basis. Expert performance is distinguished by superior results in the specific domain of interest, as in medicine. The higher performance of medical specialists has been studied in several settings, including interpretation of tests such as medical images and electrocardiograms, diagnostic interviews with simulated patients, and minimally invasive surgery (see Ericsson 2004; Ericsson 2015 for details of individual studies). The general finding is that performance improves with increasing specialization over time, as shown in Figure B.7.

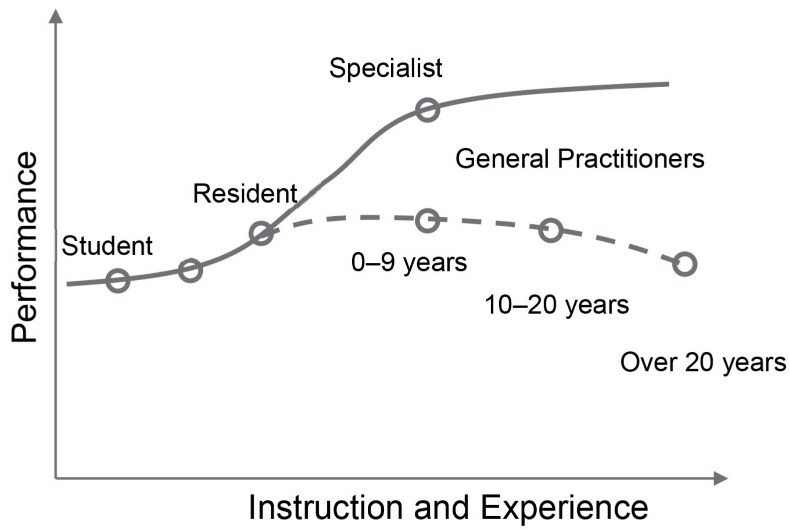

Figure B.9. Trends for development of medical performance as a function of experience and specialization based on Ericsson (2004).

In the specialties of dermatology, ECG, and microscopic pathology, diagnostic accuracy increases from a base of 20%–40% for medical students to 90% for specialists. This advantage appears to be due to better performance with more difficult cases. Mammography interpretation accuracy increases from residency to experienced specialists, and decision speed is correlated with diagnostic sensitivity for specialists but not residents. Notably, the experienced mammographers have read approximately 10,000 images in their career span, while residents might read 650 during their clinical rotation. Even among the more experienced specialists, there are substantial individual differences, and analysis of think-aloud protocols obtained during image interpretation suggests a role of self-monitoring in the application of search strategies. Interactions with patients by palliative care specialists who have received extended training in communication shows benefits in patient outcomes; however, short-duration skills training for other specialists does not show much impact, suggesting that developing expertise for patient-physician interaction is an extended process. Finally, research reviewed above suggests that specific practice on minimally invasive surgical techniques enhances the skills of both residents in training as well as more experienced surgeons.

Analysis of the training and practice environments of physicians suggests that they develop relatively standard “mental templates” for diagnosis of common illnesses. With specialization, they see more of certain kinds of illnesses, which is often associated with feedback from testing and continued interactions with patients, thus enhancing diagnostic skill. Additionally, specialists often work in clinics with more advanced equipment and knowledgeable colleagues, affording opportunities for consultation and analysis. Working in tertiary teaching environments, for example, presents circumstances for relatively continuous teaching and learning even for advanced specialists.

In popular literature, it is often stated that 10 years or 10,000 hours of experience is required to become an expert. However, it has been demonstrated (as discussed above) that the mere accrual of time is not sufficient to reach levels of superior performance, and this oft-quoted figure is a misinterpretation. In many skilled activities such as driving, acceptable performance can be achieved with around 50 hours of practice. However, this does not confer superior performance in a wide range of challenging conditions. To reach the level of expert performance, Ericsson and colleagues have proposed the theory of *deliberate practice*, which entails specific features:

- Skill exercise tasks with a well-defined goal
- Motivation to improve
- Provision of feedback
- Repetition and gradual refinements of performance.

The progression of practice and experience can lead to various levels of proficiency, as shown in Figure B.8. With everyday activities such as driving, operating computers, etc., the goal is to achieve a satisfactory level of performance relatively rapidly, in which conscious (cognitive) stages are replaced by relatively effortless or autonomous performance where it is not necessary to consciously think about individual steps. With deliberate practice, aspiring experts counteract the automaticity of performance by developing more complex mental representations; evidence for this proposition is provided by various studies of think-aloud protocols from experts in diverse domains (chess, surgery, diagnostics). In reading mammograms, for example, experts show increasing accuracy with increased time spent, whereas trainees do not (Nodine et al. 1999). In cases where the commitment to practice is reduced, performance becomes more automated and development toward expertise is arrested. Although certain curricula, such as medicine, tend to track the 10-year figure to achieve specialization, time in and of itself is insufficient.

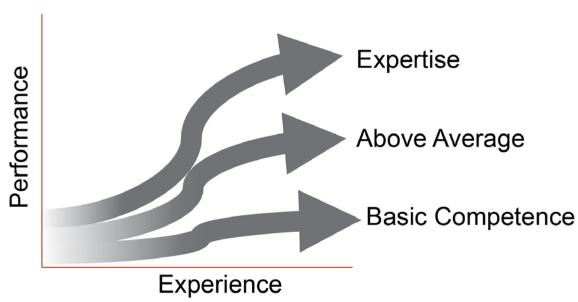

Figure B.10. Impact of practice and experience on attainment of different skill levels (Ericsson 2018).

Deliberate practice opportunities are relatively straightforward to create for entry-level personnel in many areas; learning the basic skills and procedures of a variety of crafts and professions entail the components of deliberate practice. It is more difficult to maintain this level of engagement with mid-level and experienced professionals—the general expectation is that they will reliably produce competent performance, and this is generally the case. Expert skill is most desirable for circumstances that exceed the more commonly encountered situations, such as infrequent and urgent circumstances that may arise for aircraft pilots, emergency room staff, and professionals dealing with particularly complex diagnoses or surgeries. In the case of NDE, this might entail examining a component in a highly challenging radiation environment with complex geometry and repair history that leads to ambiguities in ultrasonic signal interpretation.

Ericsson (2008) outlines ideal learning conditions for emergency physicians that embody the characteristics of deliberate practice:

- Simulations and re-creation of old cases
- Students and advanced performers are presented with cases just above their current ability level
- “Ground truth” is known, i.e., the specific corrective actions necessary are available through prior analysis of the case
- The training situation focuses on short-duration activity with immediate feedback reflection and correction
- Continuation of this type of training until consistent performance is achieved.

These general conditions are applied in varying degrees for entry-level trainees in many domains and sometimes for more advanced professional certifications, as in aircraft piloting. The utility of practicing for infrequent events was demonstrated in an analysis by McKinney Jr. and Davis (2003), who found that military pilot trainees who had practiced responding to various malfunctions in a simulator handled these same malfunctions in actual aircraft more effectively when they occurred by chance in normal operations (these were naturally occurring malfunctions, not training exercises).

Deliberate practice for practicing professionals is an area that warrants further research and development (Ericsson 2009). In medicine, standard lecture or conference-based continuing education appears to have little impact upon increasing knowledge or changing behavior, whereas interactive methods that involve coaching and feedback, reminders, prompts, etc., have a positive impact (Davis 2009). In a study of insurance agents, Sonnentag and Kleine (2000) found that frequency of various forms of deliberate practice, such as mental simulation of client interactions, asking advice from colleagues, etc., was associated with higher performance ratings for the agents. Sonnentag and Kleine (2000) point out that the structure of workplace activities can comprise supporting activities for task execution, as well as those activities which are individually undertaken to improve performance, some of which are the same. Examples include practicing presentation skills, reading the current professional literature, and collegial consultation. In those professions where deliberate practice is not built-in to the structure of work, analysis of the task activities can help identify opportunities for practice.

Finally, many professions have continuing education requirements and may also involve requalification by means of examination. Medicine, for example, has an extensive continuing medical education (CME) model, with states requiring physicians to obtain a specific number of CME hours over a several year period (ranging from 40 to 100) for license maintenance. Specialty boards require additional activities on a 10-year cycle, including CME, a knowledge examination, and assessment of practice performance for maintenance of certification (MOC) (Iglehart and Baron 2012). Simulation is used in meeting some of the MOC requirements, including anesthesiology and minimally invasive surgery (Levine et al. 2012).

B.3.8.4 Accelerating Expertise

In the routine flow of education, training, and professional advancement, expertise develops over a considerable period of time through exposure to multiple cases that challenge the skills of practitioners. The military is particularly interested in the extent to which expert proficiency might be facilitated through novel training interventions in order to produce more experts in less time. A review of this topic was conducted under the auspices of the Air Force from 2008–2009, and several demonstrations were discussed that might be considered examples of accelerated proficiency attainment (Hoffman et al. 2013). Examples in this review included two demonstrations of enhanced performance from technicians who had received intelligent tutor training and were then tested on difficult problems. Performance of the trainees equaled or exceeded that of experts who had been working in the field for over 10 years. The training regimens generally involved exposure to a high volume of technical problems in a relatively short period of time. Other examples cited in the Hoffman et al. (2013) review include the success of the Navy Top Gun training, which entailed high-volume practice in aircraft dogfighting and a variety of scenario-based military exercises using high-fidelity simulation of engagements.

The general conclusion of the Hoffman et al. (2013) review is that accelerating the development of expert proficiency is feasible, especially with the use of serious gaming and time compression to expose learners to multiple difficult problems. It should be borne in mind that these demonstrations may simply be showing the impact of overlearning on near-term retention; delayed tests were not discussed. Further, the level of resources necessary to implement accelerated proficiency learning may only be feasible in highly funded operations such as the U.S. military.

B.3.8.5 Alternatives to the Deliberate Practice Model

The foregoing discussion has reviewed research from the predominant theoretical and empirical approach to understanding expertise and practice. However, the deliberate practice theory is not the only approach, and a number of researchers suggest that while important, deliberate practice does not account for the majority of variance across individual experts (Hambrick et al. 2016). Instead, they suggest that a variety of partially heritable factors such as working memory capacity, diverse cognitive abilities, general intelligence, and personality contribute to the development of expertise, not just amount of deliberate practice. While a thorough review of these factors is beyond the scope of the current project, it is important to bear in mind that numerous variables combine with deliberate practice to result in expertise. It is noteworthy that most of the variables discussed in alternative theories are not easy to control; thus, a complex combination of genetics, personality, opportunity, parental support, and individual, deliberate practice represent the multi-factorial view of expertise. From the standpoint of regulatory guidance, practice is the only variable that is available for external influence.

B.3.8.6 Relevance for NDE Training

The effects of practice upon simple motor skills is to improve performance over time, either in terms of reducing errors, increasing speed, or both. This general principle is likely most applicable to NDE training for actions such as equipment assembly; this is a relatively routine process that can become automated with repetition over time. There can be some potential pitfalls to overlearning this skill, including paying less attention to critical instrument parameters, particularly if the equipment is unfamiliar and the examiner is executing “old scripts” for equipment assembly. However, other defenses in the overall system, such as the calibration process, should serve to identify any errors.

The more germane elements of expertise theory for NDE examiners concerns the impact of practice and experience upon development and maintenance of the cognitive skills used to identify and interpret ultrasonic signals. Research indicates that increasing specialization in professions is associated with superior performance. This suggests that the general trajectory of NDE examiner development from Level I to Level III will increase their job skills and that specific certifications such as PDI will further enhance expertise.

The research concerning deliberate practice supports the ASME Code requirement of annual practice for inspectors. However, NDE examinations do not generate a high volume of weld inspections even when working (compared to standard medical practices, for example), and 8 hours of annual practice is well below what is required for continuing education in other professions. The unspecified nature of the practice requirement leaves it open to individual interpretation, which can lead to faulty judgments of what examiners really know and need to practice. These circumstances lead to the general concern that both work volume and practice hours may be insufficient to maintain skill.

The deliberate practice research is clear on the type of practice that is most beneficial to enhancement of skill, i.e., goal-oriented, with exercises that somewhat exceed the current level of skill and are conducted in a context where coaching and feedback can be obtained. This type of practice might be best obtained for NDE examiners if practice were (1) increased to include more hours annually and (2) it was conducted according to industry-wide guidance. In our prior research, we were informed of periodic coaching sessions that were provided at EPRI, which may offer the best avenue for standardization and feedback. A further role exists for simulation, particularly if ultrasound inspection simulators can be developed that use recorded imagery files with known characteristics and which can be run on personal computers at local worksites.

B.4 Skill Retention

As discussed in Appendix A, retention of information declines over time and depends on factors such as the nature of the material, intervening activities (e.g., practice or no practice), and amount and level of original learning. This applies to learned professional skills as well, the difference being that professional skills generally integrate substantial amounts of declarative knowledge with specific decision and action sequences. The general pattern of data suggests that skill is maintained as long as practice is regular, and that age-related changes in cognition (e.g., slowing of processing speed) tend to be compensated by various means, including continued deliberate practice, the resistance of specialized skills to decline, or transitioning to different professional roles (Ericsson et al. 1993; Krampe and Charness 2018). There is agreement that continuous motor skills such as tracking and (to some extent) typing are better retained over periods of disuse than procedural skills such as instrument flight and manual landing (Schendel et al. 1978; Adams 1987; Arthur Jr. et al. 1998).

The classic study of procedural- versus perceptual-motor skill decay was conducted by Ammons et al. (1958). In this set of experiments, subjects performed two tasks: (1) a complex procedural sequence of switch manipulation, initially referring to printed guidance and over time relying on memory; and (2) a motor control task involving foot pedals and hand controls (a stick) to keep a mechanically controlled model airplane on a level course. Different groups were then re-tested at intervals ranging from 1 minute to 2 years, with no intervening practice. When the tasks were practiced to a level beyond initial proficiency, the procedural task showed proficiency loss of 50% that was regained with additional training; the perceptual-motor task showed little decay of skill even after 2 years of no practice. An example of the difference in performance decay over time between procedural and continuous tasks is shown in Figure B.11. Procedural tasks tend to involve discrete steps with multiple cognitive codes, such as verbal and visual, and also entail access to memory for branching. Perceptual-motor tasks tend to be continuous in nature because there is not a clear beginning or end, and there is a greater integration of sensory input and motor output.

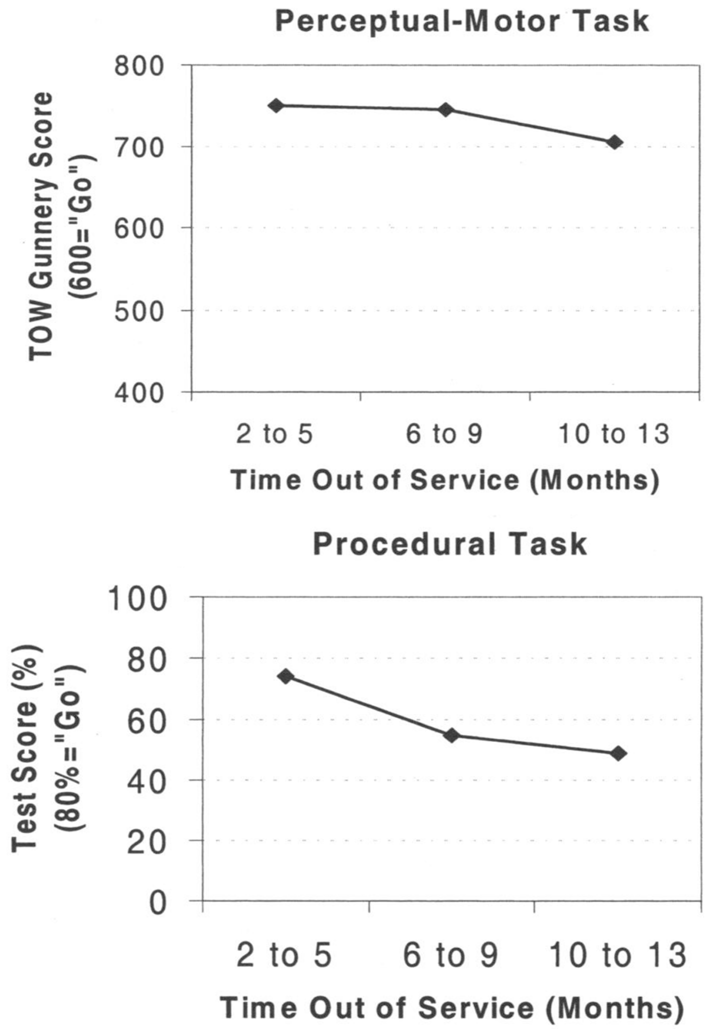

Figure B.11. Retention of continuous perceptual-motor (top) and discrete procedural tasks (bottom) after periods of disuse (Sabol and Wisher 2001). The perceptual-motor task involves marksmanship scores on gunnery tests; the procedural task involves quartermaster supply activity.

Studying professional or occupational skill retention is complicated by the relative lack of long-term periods of disuse. Individuals who undergo extensive training tend to engage in the activities for which they were trained with only occasional periods of disuse; thus, controlled experiments are rare. Instead, studies are often based on training program follow-up assessments in which the retention intervals are relatively short (e.g., less than one year), or on field studies of personnel having undergone training followed by periods of low activity, as in the military. The variety of combinations of knowledge, skill, practice, and periods of disuse are innumerable. Thus, to simplify exposition, this section reviews pertinent findings from studies of skill retention and retraining in military, aviation, and medical professions that have focused on quantifying the impact of disuse, developing models to predict levels of retention, and establishing refresher training interventions that can reduce skill decay.

B.4.1 Military

Members of the military are trained on a large variety of tasks that are not performed in everyday life, and the extent to which these tasks can be performed proficiently on-demand after no intervening practice is critical in combat situations. A number of studies were conducted in the 1970s, 1980s, and early 1990s to address this issue (e.g., Shields et al. 1979; Hagman and Rose 1983; Rose et al. 1985; Sabol and Wisher 2001).

A standard method of assessing skill levels in the military is to use the “go” rating for specific tasks; i.e., soldiers are asked to perform a specific task and expert observers determine their level of performance which results in an overall “go” or “pass” rating. Results of such a rating study for 197 volunteers from the Individual Ready Reserve are shown in Table B.4 for a series of tasks pertinent to a specific mobilization call-up. The soldiers represented in these data averaged 36 months away from active duty. It can be seen from the data that there is a wide range of performance ratings, but in no case was retention above 75% and was generally quite poor.

Table B.4. Soldier performance on common procedural tasks before retraining, following an average period of 36 months of disuse (Sabol and Wisher 2001).

Tasks	% Go
Evaluate Casualty	73
Prevent Shock	65
Apply Field Dressing	62
Wear M-17 Mask	37
Maintain M16A2 Rifle	35
Maintain M-17 Mask	34
Identify Terrain Features	28
Determine Ground Location	27
Perform Function Check M16	24
Correct Malfunction on M16	23
Determine Grid Coordinates	22
React to Chemical/Bio Hazard	19
Decontaminate Skin/Equipment	17

The variability in skill retention illustrated in Table B.4 is related to several factors that influence procedural skill retention. Studies by the Army Research Institute (ARI) over several decades have documented the following findings:

- As the number of task steps increases, retention decreases
- Retention is worse for tasks requiring recall of multiple steps and complex information

- Time limits that are difficult to meet reduce retention
- Use of job aids improves retention.

A model—the User's Decision Aid (UDA)—was developed by ARI and validated on a variety of field artillery tasks and yielded considerable predictive accuracy, particularly at a 2-month interval of disuse (the other intervals were 5 and 7 months). Correlations between predicted and actual scores were in the range of 0.9 for all retention intervals, despite the model's tendency to underestimate retention at longer intervals (Rose et al. 1985). While this model is over 30 years old, the general principles it embodies are still quite relevant when trying to assess the likelihood of retention over a prolonged period of skill disuse. The model is based on expert ratings of tasks and uses the following ten questions as a basis for rating:

1. Are job or memory aids intended to be used in this task?
2. How would you rate the quality of the job or memory aid?
3. How many steps are required to do the task?
4. Are the steps in the task required to be performed in a definite sequence?
5. Does the task have a built-in logic so you can tell if you are doing it correctly?
6. Does the task have a time limit for completion?
7. What are the mental or thinking requirements of the task?
8. How many facts, terms, names, rules, or ideas does a soldier need to memorize in order to do the task?
9. How hard are the facts, terms, etc., to remember?
10. What are the motor skill demands of the task?

Each of these questions is associated with a rating scale and guidance for application. Total scores on the task are used with look-up tables to determine the likelihood of skill decay over time and the need for retraining. The look-up tables suggest the importance of using memory or job aids for tasks that are procedural and involve numerous contingent steps; such tasks tend to be less prone to degraded performance over time because the use of memory aids can be used to refresh knowledge in real time. The UDA received continuing research attention until the early 2000s, when there was a shift in military focus. Despite the recent lack of research attention to the UDA, the underlying science is sound and could provide a basis for further development of skill decay models.

In addition to the task factors influencing skill retention described above, research with military personnel has shown that the individual (person) factor of the original training amount is also associated with retention such that greater degrees of initial learning are associated with better retention (Sabol and Wisher 2001). This finding is discussed in terms of *overlearning*, which was described in Appendix A. However, the data described by Sabol and Wisher extended to only 5 weeks beyond initial learning, so it is unclear as to the longer-term durability of overlearned tasks. The data indicating that more complex and memory-intensive tasks decay over time suggests that overlearning may be more important for showing short-term performance enhancement—as found in studies of verbal learning—than for long-term knowledge maintenance at high levels. As with studies of knowledge maintenance over decades, it is likely that more intensive and protracted training will lead to higher levels of procedural knowledge retention; in this sense, overlearning is more akin to "repeated practice" at variable intervals, similar to how students develop more advanced knowledge through higher-level course content that builds upon earlier foundational material. Studies of team performance following extended breaks also suggests that there is loss of collective skill related to team process factors such as interactivity (Cooke et al. 2013).

B.4.2 Aviation

Aviation skills comprise continuous perceptual-motor and procedural tasks and exhibit predictable declines in performance over time. An experimental study of procedural and continuous flight control skills reported by Mengelkoch et al. (1971) assessed performance in aircraft simulator exercises following periods of high and moderate proficiency training as well as a period of 4 months with no practice. The results showed that procedural tasks (e.g., cockpit check, takeoff check, low visibility procedure turn) were more susceptible to forgetting than specific aircraft control tasks as measured by deviation error from the ideal (e.g., airspeed, altitude, etc.).

An early study by the U.S. Army used a survey-estimation procedure to address skill decay following deployments and periods of proficiency flying hours with minimal flight time (Wright 1973). Ratings were obtained from 175 military aviators who had experienced periods of either minimum flying hours or extended periods of no flying. The ratings were based upon a detailed set of skills derived from analysis of flying tasks and included both continuous skills, such as visually-based flying ability, and more procedurally-oriented skills, such as developing flight plans, emergency response, etc. Ratings used an 11-point scale with ability descriptors, from 0 = unsafe to 10 = exceptional. Respondents also estimated the amount of refresher training that would take a pilot from a level of "just adequate" to "acceptable" levels of competence after periods of minimal or no flying.

The results of the survey-estimation study are complex in that many different combinations of time off and flying skill were rated, as was training necessary to refresh competence. The general findings include the following:

- Most skill loss is estimated to occur in the first 12 months of little or no flying
- Rates of skill decay decrease over time
- Visual flying skills are estimated to remain above the minimal acceptable level, while instrumented flight skills drop below the minimal acceptable level after 12 months
- Estimated training to refresh competence after 12 months of no flying is 19 hours
- Refresher training at a minimum interval of 6 months is estimated to maintain acceptable flying ability in the absence of other operational flying.

An experimental/observational study of private pilot skill loss was reported by Childs et al. (1983), in which 42 subjects were divided into three separate groups (based on the time point of interpolated training) and observed at 8, 16, and 24 months following initial certification. The observations included 27 different flight tasks and were made by senior trainer pilots using an extensive observational checklist to evaluate proficiency of trainees. The results indicated that skill loss is substantial during the first 8 months following certification (an average of 33%), with some complex tasks degrading to less than 50% correct by 24 months. Subjects were able to provide moderately accurate estimates of skill loss but not for specific flight tasks. The tasks that showed the most decline were those requiring complex cognitive procedures integrated with aircraft control, such as operations into and out of airports under adverse conditions.

The significance of the Childs et al. (1983) study is that it used in-flight observations of multiple tasks at different points in time; the results confirm the general conclusion that there is skill loss in flight tasks derived from simulator and experimental analog studies (discussed in Prophet 1976). Researchers during this time period also found that certain low-fidelity training interventions, such as “mental rehearsal” using checklists and selected images of the visual scene and aircraft instruments, was useful in mitigating the effects of skill decay. As described above, the present regulations regarding pilot flight currency requirements are well within the time window of expected skill decay.

Contemporary concerns with skill decay in aviation are more focused on the area of automation. Modern aircraft are now highly automated, and the opportunities for exercising traditional manual control and cognitive/procedural skill are correspondingly reduced. While these skills are developed to appropriate levels of proficiency during training, operational flying does not necessarily provide practice opportunities. Thus, even though operational experience may be current, various human flying skills are handled by automation, including aircraft control, navigation, and fault diagnosis. These issues were assessed in a simulator study by Casner et al. (2014), in which experienced commercial pilots were tested in conditions of flying with full, partial, and no automation. The results suggested that instrument scanning and manual control skill did not degrade, but cognitive/procedural tasks involving navigation showed operationally significant errors. Other data collected during the study suggest that automation reduces the proportion of task-relevant thoughts during flight, which may account for reduced cognitive skills. The authors suggest that more non-automated flying during operations may be appropriate, using automation as a backup, in order to exercise skills that would otherwise degrade.

B.4.3 Medical

While most studies of medical skills appear to focus on evaluating the impact of new training technologies such as simulation, the use of cardiopulmonary resuscitation (CPR) and automated external defibrillators (AEDs) have received considerable attention regarding skill decay. Both techniques are meant to respond to cardiac events (either stoppage for CPR and/or abnormal rhythms for AEDs) and require skilled application of sequential steps for proper performance. Additionally, both CPR and AED usage are uncommon; with the exception of a busy emergency room, the skills are not routinely used. Thus, CPR and AED skills provide a natural laboratory for evaluating medical skill decay, as well as training interventions to improve performance quality. It is notable that CPR is a skill for which it is relatively straightforward to identify the components of high-quality performance through task analysis, such as chest compression rate and depth and ventilation frequency. These elements can be observed by evaluators and, with modern equipment, can also be recorded in order to provide feedback and a record of performance for debriefing.

In the 1970s and 1980s, several studies suggested that CPR skills were not well-retained. Weaver et al. (1979), for example, trained a sample of 61 individuals in a 4-hour basic life support skills course using a resuscitation mannequin to measure parameters such as depth and frequency of chest compressions. Subjects were subsequently tested 6 months later, and despite initial acceptable performance, skills had declined to a point where only 11% of the trainees performed all necessary actions correctly.

Fossel et al. (1983) studied CPR skill decay in a cross-sectional design using groups tested at two weeks, one year, and two years post-training. What is notable about this study is the time-dependent decay of skill and that the 2-week group showed substantial deterioration on key component skills—adequate CPR performance was 62% in this group, 20% in the year-1 test group, and 25% in the year-2 test group (the 1-year and 2-year test groups did not differ statistically). A study by McKenna and Glendon (1985) showed a pattern of gradual decline in CPR skill over retention intervals of 2, 6, 18, and 36 months. Performance was reduced from a total correct score to 64% at 2 months and 23% at 36 months, with intermediate values for the intervening periods. Figure B.12 illustrates the skill decay trend, which is remarkably similar to forgetting curves in verbal learning and other studies of skill discussed above. These studies clearly show that a complex skill such as CPR when learned to an acceptable criterion rapidly decays to low levels without intervening practice. The loss of skill occurs across physicians, residents, nurses, and the general public (Kaye and Mancini 1986), as well as paramedics/ emergency medical technicians (Latman and Wooley 1980; Lammers et al. 2009).

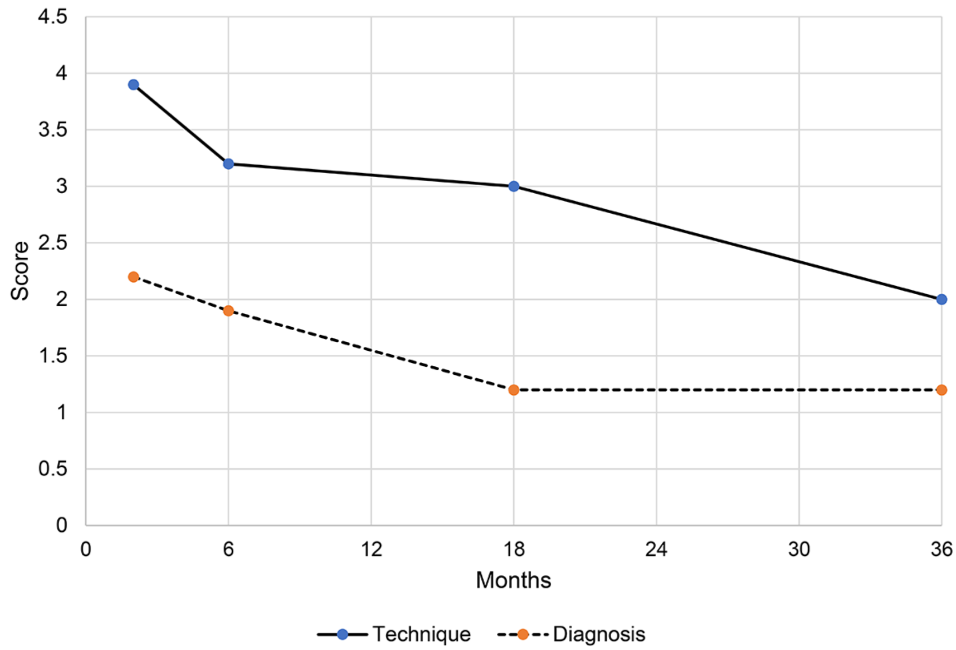

Figure B.12. Mean scores for technique and diagnosis on a CPR test. Dashed line = diagnosis (out of 4 possible), solid line = technique (out of 5 possible) (adapted from McKenna and Glendon 1985).

The studies of CPR skill decline described above were based on performance with mannequins that simulated a live patient and in some cases were able to provide quantitative measurements of performance, such as chest compression depth and rate. Live-patient studies reported by Abella et al. (2005) and Wik et al. (2005) suggest that CPR performance is suboptimal with real cardiac arrest cases by comparing measured CPR parameters to established standards. These studies were done after participating nurses and paramedics had either recently completed refresher training or were members of a team that responded to all cardiac events. The results suggest that in up to half of the case studies, chest compression was inadequate in terms of depth and frequency. These live-patient data suggest that skill decay is an issue with specific clinical consequences. There was also a suggestion of disparity between mannequin studies and live-patient data, such that mannequin studies tend to find chest compressions that are too deep, while the opposite occurs in live patients.

Several approaches have been investigated to improve the quality of CPR delivered in actual patient care, based on the traditional learning principles of feedback, more frequent spaced training, and debriefing following resuscitations. Feedback can be provided either by the

CPR/defibrillation device using visual and auditory cues (e.g., “deeper,” “faster,” etc.) or by subsequently reviewing recordings of individual cases made by the device. Meta-analytic review of mannequin and human intervention studies using real-time feedback (Kirkbright et al. 2014) suggest that CPR quality is improved with feedback such that chest compression depth and other measures are improved. Bobrow et al. (2013) reported a field study of out-of-hospital cardiac arrest CPR in which a “bundle” approach to training was provided, which included training sessions specifically focused on CPR quality parameters and the use of defibrillator audio-visual feedback on patients resuscitated post-training. The results indicated that the training intervention and subsequent use of feedback during actual resuscitation events improved patient outcomes such that there was a significant increase in the percentage of patients surviving to hospital discharge (5.2% more). Further, all parameters of CPR quality measured by the defibrillator in the post-training period showed significant improvement. A similar training bundle study for in-patient cardiac arrest resuscitations was reported by Davis et al. (2015), which showed that survival to discharge was improved and increased for several years during the intervention period. The Davis et al. (2015) study appeared to involve relatively continuous monitoring and feedback/debriefing to resuscitation team members, rather than being a single training session followed by device-based feedback as in the Bobrow et al. (2013) study. Providing training at the site of anticipated resuscitations, such as a pediatric intensive care unit, has also been shown to improve clinical team confidence, particularly if it was given more than two times per month (Niles et al. 2009).

B.4.4 Relevance for NDE Training

The rapid decay of procedural skill during periods of disuse occurs across diverse performance contexts, including the military, aviation, and medicine. The decay functions are remarkably similar, with substantial loss occurring in the first several weeks or months and, depending on the specific task, often falling below 50% retention. This loss of skill occurs for seasoned professionals as well as more junior personnel; for complex procedure-oriented skill, the adage “use it or lose it” seems to apply. Much less is known about optimal refresher training frequency—such studies are logistically very difficult and retraining and certification processes tend to be set by professional licensing or certification authorities. The general recommendation from skill decay studies is that training be provided, if feasible, within the window of steep skill decline, which appears to be the first 3 months of disuse. The medical studies of “rolling refreshers,” in which mannequin-based CPR training is provided in the intensive care units, is similar to use of “just-in-time” training that provides the opportunity to practice procedures prior to operational inspection.

NDE inspection in the nuclear industry is associated with regular periods of disuse. The concentration of inspections during fall and spring outages can lead to periods as long as 5 or 6 months between application of the relevant skills. The ASME Code requires 8 hours of annual training, although the content and scheduling of that training is unspecified. Our prior research (Sanquist et al. 2018) suggests that NDE examiners individually practice with samples available to them, particularly prior to engaging in inspections during an upcoming outage season. Since studies of multiple types of professional skill suggest steep decay in the first 3 to 6 months of disuse, it is reasonable to conclude that the same occurs for NDE skills.

The finding that procedural skills decline, whereas continuous perceptual-motor skills are more durable, suggests that NDE practice should be focused on more complex procedures rather than on simpler skills that may be well-established, such as equipment assembly, establishing transducer coupling, and moving/skewing the transducer probe. A likely area of skill degradation would be the use of the EPRI flowchart for evaluating indications observed during scanning—

these are multi-page logic sequences with branches that depend on additional findings. Use of these flowcharts is likely to be less frequent than the governing PDI procedure, since indications are not observed during every use of a particular procedure. Similarly, decay of memory for steps within an “informational use” procedure has been shown to be related to specific inspection errors (Sanquist et al. 2018).

The skill decay literature illustrates the utility of task analysis; in all domains reviewed—military, aviation, and medical—task analytic decomposition of procedural skill components was used to characterize the task elements that decay over time. Further, the aviation studies suggest that use of behavioral anchors on rating scales is a useful technique for estimating decay and training needs to become proficient again. There may be existing sources of data, primarily with EPRI, that can be used to determine more precisely which procedures are the most likely to show the effects of disuse over time; the military studies of the UDA suggest that those procedures involving more steps, fewer memory aids, and more complex sequences would be the most prone to skill decay.

B.5 Summary and Principles for Application of Research and Theory

This appendix has reviewed research from the science of training and instruction as it pertains to the core issues in NDE training—amount of time spent in the field to develop skill, the progressive loss of skill over periods of disuse, and the importance of specific training and instructional factors such as practice, feedback, active and observational learning, visualization and multimedia, and simulation. In this summary section, we review the key points of prior sections in the context of what they might tell us about implementing NDE training and articulate core principles of training and instruction derived from the studies reviewed.

The first question addressed in this review concerned the extent to which findings from laboratory studies of learning and memory are generalizable to application settings. Two very consistent findings in laboratory studies are (1) the superior retention of material that is studied in a spaced rather than massed fashion and (2) the facilitative effects of knowledge testing. The research reviewed showed a number of demonstration studies conducted with medical students and illustrated that both study spacing and knowledge testing enhance retention beyond what is obtained with the same study time that is massed or simply additional study time. They key principle resulting from this research is that **distributed (spaced) practice of occupationally relevant material and knowledge testing of that same material yields superior retention to simple re-study**. The beneficial effects in these studies were obtained on the basis of planning for the presentation or testing of curriculum material in such a way as to yield spaced presentation or knowledge testing.

The review identified very few articles that specifically focused on the topic of OJT. Distinctions were made between on-the-job experience and on-the-job training. The latter is considered to be a structured process that is overseen by more senior professionals; simple “experience” in the job setting does not necessarily lead to specific knowledge acquisition and ability to apply it. A prescriptive model of OJT described by Semb et al. (2000) discussed OJT in terms of the primary elements of instructional system design, i.e., assessment, instruction, and evaluation. Assessment is based on task analysis of the job to be performed and identifies the specific technical knowledge that a learner needs to master. Instruction in the OJT environment is similar to tutoring or coaching where the senior professional develops instructional opportunities based on operational circumstances. In its most elaborated form, OJT instruction employs many different techniques that have been shown to be effective in teaching, such as incremental instruction, shaping of responses, feedback, etc. Evaluation in the OJT context is often linked to instruction, with gradual increases in task responsibility by the trainee and the instructor appraising performance in operational tasks. The key principle for OJT is that **OJT is a systematic process based on analysis of the task to be performed, development of instructional situations from the operational environment, employment of instructional techniques as warranted, and evaluation of performance based on increasing competence in operational tasks.** To the extent that certification or regulatory processes specify tasks or operational circumstances to be completed, as in aviation, the more likely field experience requirements are to be structured to accomplish them.

In the training and instructional psychology literature, feedback is considered in the context of the overall learning process, as distinct from the more test-item-specific feedback addressed in laboratory studies. The instructional process entails an extended set of interactions between learner and instructor and provides opportunities for feedback at multiple levels. Feedback theory has identified four levels at which learners can be provided with performance information: task, process, self-regulation, and self. The task level is addressed with feedback concerning specific performance on test items—correct or incorrect, or correct answer. Task feedback is more effective when it is specific, such as individual item responses, rather than general, such as overall test performance. Process feedback is concerned with what learners do in order to accomplish a task, including their method of problem solving, strategies, concepts used, etc. This level of feedback is aimed at improving task performance by enhancing the process. Self-regulation involves the learner developing the ability to monitor and appraise their own performance; it is generated internally but is based on prior task and process-level feedback. Feedback concerning "self" tends to involve instructor praise that is not task or process-specific, and thus not particularly effective. A key principle from the review is that **feedback is a multi-layered process addressing task and process elements that can lead to internal cognitive structures for self-monitoring**.

Further work in the area of feedback addressing delayed instruction, error management, scenario reviews and the use of supportive materials (learning criteria, checklists, reflection sheets) suggests that **feedback is a constructive process which can actively engage the learner**. This active engagement facilitates development of self-monitoring. **Feedback is thus a continuous process of constructive evaluation that can become part of ongoing operations**.

The constructive view of feedback is incorporated across the spectrum of learning types and instructional techniques in the theory of active learning. A variety of research findings indicate that as students engage with material in more active ways than being passive recipients of information, their comprehension, retention, and ability to apply knowledge improves. This pattern of findings has been described as ICAP (Interactive, Constructive, Active, Passive), to characterize learning modes with increasing levels of student engagement both with study material and with each other. Increases in activity level, for example from passive to active, may simply involve an instructional manipulation such as highlighting material and reviewing it rather than just reading. Constructive and interactive modes entail more elaborate interactions with the material, including conceptual mapping, generating self-explanations, and group discussions about material. The general principle illustrated by this work is that **as learning modes involve increasing level of material and student engagement, overall benefits in terms of knowledge structure, retention, and application are improved.** Active learning modes can involve a variety of instructional techniques, including studying worked examples and generating self-explanations. More traditional approaches such as lecture and discussion can also be adapted to incorporate higher levels of student activity.

Observation is a core element of all learning processes, whether in a classroom setting or on the job. In the context of instructional techniques, observational learning is specifically defined as watching a role model carry out target behaviors (behavior modeling) or parts of the job task (DBT). In either case, the learner sees examples of expected performance and can be provided with the opportunity to reproduce what they have seen. Observational learning involving full BMT would entail watching a scripted sequence of actions, whereas DBT would involve selecting specific elements of a job to illustrate. Both approaches complement other training techniques and provide concrete examples of expected performance. The key principle of observational learning is that **student observation of whole or part performance can be flexibly implemented in training or operational settings to illustrate integration of knowledge.**

The multimedia instruction hypothesis—that learning is better with words and pictures than with words alone—has been confirmed in numerous experiments. The superior effect of multimedia is related to the structure of the human cognitive system, which encodes information in separate channels most appropriate to the sensory stream. Research has elucidated a number of design principles of visualization and multimedia that serve to focus attention through information selection, enhance essential information processing through organization, and integrate the material with other knowledge. Visualization and multimedia development capabilities are widely available, and it is important to adhere to design principles to reduce excessive material. The core principle from studies of visualization and multimedia is that **learning is enhanced by using multiple sensory modes of presentation for material that is designed to select, organize, and integrate learning content**. Studies of learning conducted with material presented by intuitive instructor design versus the same material redesigned according to multimedia principles shows that learning and retention are improved.

Simulators are widely applied in many training programs, particularly aviation. The core issue from the standpoint of personnel qualification is the extent to which a simulator can be used to substitute for training in the field, whether it be operational aircraft, clinically supervised surgeries, or supervised experience during in-service NDE. This is known as “syllabus reduction,” i.e., the percentage reduction in operational training time that might be achieved with simulation. In aviation, studies over the past 60 years have reported an extremely wide range (1%–93%). More recent studies suggest reductions in operational aircraft time of 25%. Regulation adopts a conservative stance—for pilot certification, no more than 100 of the required 1500 hours of flight time may be obtained in a simulator, with more liberal substitution for approved training programs and various ratings once a pilot is certified. Studies of surgical simulation suggest that there is positive transfer from simulation to clinical experience, with the potential to reduce training time by as much as 40%. The general principle resulting from this work is that **simulation results in positive transfer of training and can be used to reduce operational experience requirements**. The literature reviewed also suggests that determining a fixed percentage of syllabus reduction and simulator transfer effectiveness is experimentally intractable due to the many different combinations of simulator vs. operational time required and that **for high-risk jobs, simulation should be considered an “accelerator” rather than a substitute.**

The predominant theoretical view of developing expertise is that of deliberate practice. Studies of simple skills indicate that continuing practice yields improvement over periods of years. Theory and supporting data suggest that proficiency at professional tasks develops through a series of competency levels, with most average practitioners achieving acceptable proficiency at the level of automated skill and improving no further despite long tenure in a profession. Experts are defined as practitioners who reliably achieve superior results. When opportunities for practitioners are provided to specialize and perform many more hours of work in their domain, expertise can be developed. Specific numbers of hours necessary to achieve expertise is elusive, although popular accounts have misinterpreted research to suggest that there is a 10,000-hour minimum. While 10,000 hours is a guide, it is not a hard requirement; instead, it reflects the progression over a time period in which opportunities for practice are available. The general principle that results from this work is that **practice in high volume upon tasks that challenge current skill in association with motivation and feedback can lead to higher levels of proficiency**.

The skill retention literature illustrates that skill decay is a predictable consequence across numerous types of tasks, both laboratory learning experiments as discussed in Appendix A and diverse application areas including military, aviation, and medical domains. Continuous perceptual-motor tasks—akin to vehicle control—tend to be preserved over periods of disuse, whereas tasks involving complex cognitive and procedural execution with various decision points and contingencies tend to degrade relatively quickly, often within the first several months of disuse. Evidence suggests that skill decay is somewhat mitigated by a higher level of original learning, but it is not eliminated. Parametric studies of refresher training have not been conducted, but available evidence suggests that skill levels can be re-established with training provided during the period of disuse even through such simple means as mental rehearsal. The general principle associated with this research is that **cognitive-procedural skills decline rapidly during periods of disuse and can be mitigated by training provided during the period when skills are not being exercised in an operational environment**.

With respect to NDE training, the foregoing principles suggest several potential applications across the spectrum of initial fundamental training, field experience, and skill maintenance during down-time:

- To the extent practicable, incorporating the concepts of distributed (spaced) practice and knowledge testing during course work that conveys foundational material will enhance retention of that knowledge.
- Presentation of foundational learning material through visualization and multimedia designed according to psychological principles can enhance attention to relevant material and integration with existing knowledge.
- Use of active-learning techniques in foundational training as well as operational field experience will enhance learning and retention.
- Field experience provides the opportunity for structured OJT and should be provided to junior personnel in a way that avoids the simple accrual of time in the field. Instead, specifically structured activities that provide the opportunity to execute increasingly responsible tasks under the supervision of senior personnel is the best way to implement field experience. OJT should incorporate behavior modeling and DBT as appropriate to the specific circumstances.
- Feedback across both the training and operational experiences should be routinely provided both in terms of specific task outcomes and the processes used by individuals. Active engagement of feedback can help to develop self-monitoring mechanisms and use of post-inspection debriefings can enhance the inspection process both from the standpoint of the individual examiner and the utility.
- Laboratory exercises and/or ultrasonic inspection simulation will yield transfer of training and can be a learning accelerator, but a conservative approach to substitution for operational experience is suggested by practices in other industries.

- Continued practice both through operational inspections and use of practice samples will improve skill levels. Expertise requires many years of experience to develop, particularly in circumstances where difficult cases are rare. This suggests that development of specific practice samples and scenarios that incorporate the characteristics of difficult cases that have been encountered would be a useful addition to the overall training and practice experience of examiners.
- Cognitive-procedural skill will decay during periods of disuse, and training and practice during these down-times is warranted to maintain proficiency.

These suggestions for application of principles derived from the instructional and training literature can provide impetus for review of existing NDE training and field experience practices and potentially identify opportunities for implementation.

B.5.1.1 Implications of the Psychology of Learning, Memory, Training, and Instruction for ASME Code Technical Basis

The review of scientific literature documented in this appendix was conducted to determine the extent to which there exists a basis for reducing the number of experience hours to attain an initial Level II rating and to identify material pertinent to skill maintenance via annual practice. The industry-proposed change in field experience requirements reduces the number of hours from 800 to 400, with 320 of those hours permissible as laboratory training. The overall reduction of in-field experience is thus as high as 90%, with lab or simulated flaw training substitution of 80%. No change to annual practice requirements of 8 hours is proposed.

The key findings from the literature review indicate that **there is no technical basis for reducing field experience time requirements**; greater amounts of learning time lead to greater levels of knowledge and retention. Knowledge and procedural skill decay very quickly during periods of disuse, and the rate of decay is similar for experts and non-experts. Knowledge decays from the overall level attained, so there will be lower amounts of loss for those individuals with higher levels of knowledge; complex procedures that are memory-intensive and involve multiple branching steps are particularly prone to skill decay. Studies of simulator substitution and the regulations governing the airline industry suggest that simulator training is acceptable for small substitution percentages during initial training and higher percentages for specific aircraft types and certifications (e.g., instrumented flight) once the primary certification is obtained. As an example, initial commercial pilot certification requires 1500 hours of operational aircraft flight experience, for which 100 hours in an approved simulator may be substituted. In medicine, simulation is required in various surgical curricula but does not substitute for clinical experience.

There are certain practical considerations associated with field experience that warrant consideration. The most significant consideration is the low rate of occurrence of flaws in operational settings. Detecting and characterizing flaws and indications is the most critical part of the examiner's job, but is not necessarily exercised in a way that will allow for skill development. Thus, more laboratory practice would provide this opportunity. However, reduction to 80 hours of operational experience would be excessive—there are numerous intangible elements of the examiner job that are learned through field experience, including:

- learning how to navigate through containment/inspection areas
- executing procedures in complex circumstances
- observing multiple geometries

- addressing issues of coverage limitations
- generally learning flexibility to adapt the examination process while maintaining compliance with the qualified procedure
- experience in diverse operational cultures and expectations.

For new personnel with little to no experience in the field of NDE, the 800-hour requirement is warranted; 320 hours of laboratory training will increase opportunities to practice detection and sizing of flaws by use of laboratory practice exercises that simulate the types of examinations that would be done in the field; the balance of 480 hours of operational experience will permit the development of practical skills in realistic settings. **For NDE examiners who already have considerable experience in other special industry sectors, it may be appropriate to reduce the 800-hour requirement and substitute examination of practice samples to build knowledge of special nuclear industry materials and flaw characteristics. This would be coupled with appropriate PDI qualification**.

With respect to practice and experience once initial knowledge is established, the overriding conclusion is similar to that of initial learning—more practice and experience leads to greater levels of expertise. Retraining after periods of time off has not been examined parametrically in the literature except in artificial laboratory verbal learning studies. **It is thus not possible to determine whether the currently required 8 hours of practice on operationally-similar specimens is sufficient.** Laboratory data suggest that less time is required to regain mastery after skill or knowledge decays (there is a “savings” in re-learning), but the extent to which such savings translate to operational environments is unclear. Certain skilled professions such as medicine require continuing education annually of 40 hours or more; for pilots who have not flown for more than 90 days, three takeoffs and landings in an approved visual simulator must be performed. Studies of CPR skill suggest that continuing training and feedback embedded into the operational environment on a case-by-case basis is the best approach to skill maintenance. Because testing has been shown to enhance retention of existing knowledge, it may be appropriate to develop further guidance on how to structure and test examiner practice on relevant samples, as there is currently no structure specified in the Code. The apparently self-directed nature of annual practice is not optimal from the standpoint of self-monitoring of knowledge; the literature suggests that people are not particularly good at judging what they should study and how much they should study. Thus, some **more specific structure and guidance of annual practice that includes studying material pertinent to upcoming examination and knowledge testing of what is studied would be appropriate.**

B.6 References

Abella BS, JP Alvarado, H Myklebust, DP Edelson, A Barry, N O’Hearn, TL Vanden Hoek and LB Becker. 2005. "Quality of Cardiopulmonary Resuscitation During In-Hospital Cardiac Arrest." *JAMA* 293(3):305-310. DOI: 10.1001/jama.293.3.305.

Adams JA. 1979. "On the Evaluation of Training Devices." *Human Factors* 21(6):711-720. DOI: 10.1177/001872087912210608.

Adams JA. 1987. "Historical Review and Appraisal of Research on the Learning, Retention, and Transfer of Human Motor Skills." *Psychological Bulletin* 101(1):41-74. DOI: 10.1037/0033-2909.101.1.41.

Adams PS, RL Brauer, B Karas, TF Bresnahan and H Murphy. 2004. "Professional Certification: Its Value to SH&E Practitioners and the Profession." *Professional Safety* 49(12):26-31.

Ammons RB, RG Farr, E Bloch, E Neumann, M Dey, R Marion and CH Ammons. 1958. "Long-Term Retention of Perceptual-Motor Skills." *Journal of Experimental Psychology* 55(4):318-328. DOI: 10.1037/h0041893.

Arthur Jr. W, W Bennett Jr., PL Stanush and TL McNelly. 1998. "Factors That Influence Skill Decay and Retention: A Quantitative Review and Analysis." *Human Performance* 11(1):57-101. DOI: 10.1207/s15327043hup1101_3.

Ausubel DP. 1960. "The Use of Advance Organizers in the Learning and Retention of Meaningful Verbal Material." *Journal of Educational Psychology* 51(5):267-272. DOI: 10.1037/h0046669.

Baghdady M, H Carnahan, EWN Lam and NN Woods. 2014. "Test-Enhanced Learning and Its Effect on Comprehension and Diagnostic Accuracy." *Medical Education* 48(2):181-188. DOI: 10.1111/medu.12302.

Bahrick HP, LE Bahrick, AS Bahrick and PE Bahrick. 1993. "Maintenance of Foreign Language Vocabulary and the Spacing Effect." *Psychological Science* 4(5):316-321. DOI: 10.1111/j.1467-9280.1993.tb00571.x.

Bailey T. 1993. "Can Youth Apprenticeship Thrive in the United States?" *Educational Researcher* 22(3):4-10. DOI: 10.3102/0013189x022003004.

Bandura A. 1977. *Social Learning Theory*, Englewood Cliffs, NJ: Prentice Hall. 247. ISBN 978-0138167516.

Bangert-Drowns RL, JA Kulik and C-LC Kulik. 1991. "Effects of Frequent Classroom Testing." *The Journal of Educational Research* 85(2):89-99. DOI: 10.1080/00220671.1991.10702818.

Bell BS and SWJ Kozlowski. 2009. "Toward a Theory of Learner-Centered Training Design: An Integrative Framework of Active Learning." In *Learning, Training, and Development in Organizations*, pp. 263-300, eds: SWJ Kozlowski and E Salas. New York: Routledge.

Bell BS, SI Tannenbaum, JK Ford, RA Noe and K Kraiger. 2017. "100 Years of Training and Development Research: What We Know and Where We Should Go." *Journal of Applied Psychology* 102(3):305-323. DOI: 10.1037/apl0000142.

Blum T, A Rieger, N Navab, H Friess and M Martignoni. 2013. "A Review of Computer-Based Simulators for Ultrasound Training." *Simulation in Healthcare* 8(2):98-108. DOI: 10.1097/SIH.0b013e31827ac273.

Bobrow BJ, TF Vadeboncoeur, U Stolz, AE Silver, JM Tobin, SA Crawford, TK Mason, J Schirmer, GA Smith and DW Spaite. 2013. "The Influence of Scenario-Based Training and Real-Time Audiovisual Feedback on Out-of-Hospital Cardiopulmonary Resuscitation Quality and Survival From Out-of-Hospital Cardiac Arrest." *Annals of Emergency Medicine* 62(1):47-56.e1. DOI: 10.1016/j.annemergmed.2012.12.020.

Breymier TL, T Rutherford-Hemming, TL Horsley, T Atz, LG Smith, D Badowski and K Connor. 2015. "Substitution of Clinical Experience With Simulation in Prelicensure Nursing Programs: A National Survey in the United States." *Clinical Simulation in Nursing* 11(11):472-478. DOI: 10.1016/j.ecns.2015.09.004.

Broadwell MM. 1986. *Supervising Today: A Guide for Positive Leadership*, New York: Wiley-Interscience. ISBN 978-0471836742.

Brown AL. 1975. "The Development of Memory: Knowing, Knowing about Knowing, and Knowing How to Know." In *Advances in Child Development and Behavior, Vol. 10*, pp. 103-152, ed: HW Reese. New York: Academic Press.

Bryan WL and N Harter. 1899. "Studies on the Telegraphic Language: The Acquisition of a Hierarchy of Habits." *Psychological Review* 6(4):345-375. DOI: 10.1037/h0073117.

Butler AC, EJ Marsh, JP Slavinsky and RG Baraniuk. 2014. "Integrating Cognitive Science and Technology Improves Learning in a STEM Classroom." *Educational Psychology Review* 26(2):331-340. DOI: 10.1007/s10648-014-9256-4.

Caldwell JE. 2007. "Clickers in the Large Classroom: Current Research and Best-Practice Tips." *CBE—Life Sciences Education* 6(1):9-20. DOI: 10.1187/cbe.06-12-0205.

Campbell JP. 1971. "Personnel Training and Development." *Annual Review of Psychology* 22(1):565-602. DOI: 10.1146/annurev.ps.22.020171.003025.

Cannon-Bowers JA, E Salas, SI Tannenbaum and JE Mathieu. 1995. "Toward Theoretically Based Principles of Training Effectiveness: A Model and Initial Empirical Investigation." *Military Psychology* 7(3):141-164. DOI: 10.1207/s15327876mp0703_1.

Caro PW. 1973. "Aircraft Simulators and Pilot Training." *Human Factors* 15(6):502-509. DOI: 10.1177/001872087301500602.

Carpenter SK. 2017. "Spacing Effects on Learning and Memory." In *Cognitive Psychology of Memory, Vol. 2 of Learning and Memory: A Comprehensive Reference, 2nd Edition*, pp. 465-485, ed: JT Wixted. Oxford: Academic Press.

Carretta TR and RD Dunlap. 1998. *Transfer of Training Effectiveness in Flight Simulation: 1986 to 1997*. AFRL-HE-AZ-TR-1998-0078. Mesa, AZ: Air Force Research Laboratory.

Casner SM, RW Geven, MP Recker and JW Schooler. 2014. "The Retention of Manual Flying Skills in the Automated Cockpit." *Human Factors* 56(8):1506-1516.

Cathcart JA and G Graff. 1978. "Occupational Licensing: Factoring It Out." *Pacific Law Journal* 9:147-164.

Chi MTH. 2009. "Active-Constructive-Interactive: A Conceptual Framework for Differentiating Learning Activities." *Topics in Cognitive Science* 1(1):73-105. DOI: 10.1111/j.1756-8765.2008.01005.x.

Chi MTH, M Bassok, MW Lewis, P Reimann and R Glaser. 1989. "Self-explanations: How Students Study and Use Examples in Learning to Solve Problems." *Cognitive Science* 13(2):145-182. DOI: 10.1016/0364-0213(89)90002-5.

Chi MTH and R Wylie. 2014. "The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes." *Educational Psychologist* 49(4):219-243. DOI: 10.1080/00461520.2014.965823.

Childs JM, WD Spears and WW Prophet. 1983. *Private Pilot Flight Skill Retention 8, 16, and 24 Months Following Certification*. DOT/FAA/CT-83/34; SEVILLE TR-83-17. Atlantic City Airport, NJ: Federal Aviation Administration.

Cooke N, J Gorman, C., J Duran, CW Myers and DH Andrews. 2013. "Retention of Team Coordination Skill." In *Individual and Team Skill Decay*, pp. 344-363, eds: W Arthur Jr., EA Day, W Bennett Jr. and AM Portrey. NY: Routledge.

Craik FIM and E Tulving. 1975. "Depth of Processing and the Retention of Words in Episodic Memory." *Journal of Experimental Psychology: General* 104(3):268-294. DOI: 10.1037/0096-3445.104.3.268.

Crossman ERFW. 1959. "A Theory of the Acquisition of Speed-skill." *Ergonomics* 2:153-166. DOI: 10.1080/00140135908930419.

Davis DA. 2009. "How to Help Professionals Maintain and Improve Their Knowledge and Skills: Triangulating Best Practices in Medicine." In *Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments*, pp. 180-202. New York, NY: Cambridge University Press.

Davis DP, PG Graham, RD Husa, B Lawrence, A Minokadeh, K Altieri and RE Sell. 2015. "A Performance Improvement-Based Resuscitation Programme Reduces Arrest Incidence and Increases Survival from In-Hospital Cardiac Arrest." *Resuscitation* 92:63-69. DOI: 10.1016/j.resuscitation.2015.04.008.

Dawe SR, GN Pena, JA Windsor, JAJL Broeders, PC Cregan, PJ Hewett and GJ Maddern. 2014. "Systematic Review of Skills Transfer After Surgical Simulation-Based Training." *British Journal of Surgery* 101(9):1063-1076. DOI: 10.1002/bjs.9482.

Diehl AE and LE Ryan. 1977. *Current Simulator Substitution Practices in Flight Training*. TAEG Report No. 43. Orlando, FL: Training Analysis and Evaluation Group.

Dyre L, A Tabor, C Ringsted and MG Tolsgaard. 2017. "Imperfect Practice Makes Perfect: Error Management Training Improves Transfer of Learning." *Medical Education* 51(2):196-206. DOI: 10.1111/medu.13208.

Ellis S and I Davidi. 2005. "After-Event Reviews: Drawing Lessons From Successful and Failed Experience." *Journal of Applied Psychology* 90(5):857-871. DOI: 10.1037/0021-9010.90.5.857.

EPRI. 2008. *Nondestructive Evaluation: Fast-Track NDE Work Force Enhancement*. Report 10016670. Palo Alto, CA: Electric Power Research Institute.

Ericsson KA. 2004. "Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains." *Academic Medicine* 79(10):S70-S81.

Ericsson KA. 2008. "Deliberate Practice and Acquisition of Expert Performance: A General Overview." *Academic Emergency Medicine* 15(11):988-994. DOI: 10.1111/j.1553-2712.2008.00227.x.

Ericsson KA. 2009. "Enhancing the Development of Professional Performance: Implications from the Study of Deliberate Practice." In *Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments*, pp. 405-431. New York, NY: Cambridge University Press.

Ericsson KA. 2015. "Acquisition and Maintenance of Medical Expertise: A Perspective From the Expert-Performance Approach With Deliberate Practice." *Academic Medicine* 90(11):1471-1486. DOI: 10.1097/acm.0000000000000939.

Ericsson KA. 2018. "The Differential Influence of Experience, Practice, and Deliberate Practice on the Development of Superior Individual Performance of Experts." In *The Cambridge Handbook of Expertise and Expert Performance*, pp. 745-769, 2 ed. eds: AM Williams, A Kozbelt, KA Ericsson and RR Hoffman. Cambridge: Cambridge University Press.

Ericsson KA, RT Krampe and C Tesch-Römer. 1993. "The Role of Deliberate Practice in the Acquisition of Expert Performance." *Psychological Review* 100(3):363-406. DOI: 10.1037/0033-295X.100.3.363.

Ericsson KA and AC Lehmann. 1996. "Expert and Exceptional Performance: Evidence of Maximal Adaptation to Task Constraints." *Annual Review of Psychology* 47(1):273-305. DOI: 10.1146/annurev.psych.47.1.273.

Fitts PM and MI Posner. 1967. *Human Performance*, Oxford, England: Brooks/Cole.

Fossel M, RT Kiskaddon and GL Sternbach. 1983. "Retention of Cardiopulmonary Resuscitation Skills by Medical Students." *Journal of Medical Education* 58(7):568-575.

Freeman S, SL Eddy, M McDonough, MK Smith, N Okoroafor, H Jordt and MP Wenderoth. 2014. "Active Learning Increases Student Performance in Science, Engineering, and Mathematics." *Proceedings of the National Academy of Sciences* 111(23):8410-8415. DOI: 10.1073/pnas.1319030111.

Gagne RM, LJ Briggs and WW Wager. 1992. *Principles of Instructional Design, 4th Edition*, NY: Holt, Rinehart and Winston. ISBN 9780030347573.

Gagne RM and WD Rohwer Jr. 1969. "Instructional Psychology." *Annual Review of Psychology* 20(1):381-418. DOI: 10.1146/annurev.ps.20.020169.002121.

Gagné RM and EC Smith Jr. 1962. "A Study of the Effects of Verbalization on Problem Solving." *Journal of Experimental Psychology* 63(1):12-18. DOI: 10.1037/h0048703.

Gallagher AG, NE Seymour, J-A Jordan-Black, BP Bunting, K McGlade and RM Satava. 2013. "Prospective, Randomized Assessment of Transfer of Training (ToT) and Transfer Effectiveness

Ratio (TER) of Virtual Reality Simulation Training for Laparoscopic Skill Acquisition." *Annals of Surgery* 257(6):1025-1031. DOI: 10.1097/SLA.0b013e318284f658.

Gates AI. 1917. "Recitation as a Factor in Memorizing." *Archives of Psychology* 6(40):140.

Goldstein IL. 1980. "Training in Work Organizations." *Annual Review of Psychology* 31(1):229-272. DOI: 10.1146/annurev.ps.31.020180.001305.

Graesser AC. 1993. *Questioning Mechanisms During Tutoring, Conversation, and Human-Computer Interaction*. ONR/TR 93-1. Arlington, VA: Office of Naval Research.

Gray WD. 2017. "Plateaus and Asymptotes:Spurious and Real Limits in Human Performance." *Current Directions in Psychological Science* 26(1):59-67. DOI: 10.1177/0963721416672904.

Greenfield PM. 1984. "A Theory of the Teacher in the Learning Activities of Everyday Life." In *Everyday Cognition: Its Development in Social Context*, pp. 117-138, eds: B Rogoff and J Lave. Cambridge, MA: Harvard University Press.

Greenwood CR, JC Delquadri and RV Hall. 1984. "Opportunity to Respond and Student Academic Performance." In *Behavior Analysis in Education*, pp. 58-88, eds: W Heward, T Heron, D Hill and J Trap-Porter. Columbus, OH: Charles E. Merrill.

Hagman JD and AM Rose. 1983. "Retention of Military Tasks: A Review." *Human Factors* 25(2):199-213.

Hambrick DZ, BN Macnamara, G Campitelli, F Ullén and MA Mosing. 2016. "Chapter One - Beyond Born versus Made: A New Look at Expertise." *Psychology of Learning and Motivation* 64:1-55. DOI: 10.1016/bs.plm.2015.09.001.

Hattie J, M Gan and C Brooks. 2017. "Instruction Based on Feedback." In *Handbook of Research on Learning and Instruction, 2nd Edition*, pp. 290-324, eds: RE Mayer and PA Alexander. NY: Routledge.

Hattie J and H Timperley. 2007. "The Power of Feedback." *Review of Educational Research* 77(1):81-112. DOI: 10.3102/003465430298487.

Hoffman RR, P Ward, PJ Feltovich, L DiBello, SM Fiore and DH Andrews. 2013. *Accelerated Expertise: Training for High Proficiency in a Complex World*, New York, NY: Taylor and Francis. ISBN 978-1848726512.

Iglehart JK and RB Baron. 2012. "Ensuring Physicians' Competence — Is Maintenance of Certification the Answer?" *New England Journal of Medicine* 367(26):2543-2549. DOI: 10.1056/NEJMhpr1211043.

Issa N, RE Mayer, M Schuller, E Wang, MB Shapiro and DA DaRosa. 2013. "Teaching for Understanding in Medical Classrooms Using Multimedia Design Principles." *Medical Education* 47(4):388-396. DOI: 10.1111/medu.12127.

Issa N, M Schuller, S Santacaterina, M Shapiro, E Wang, RE Mayer and DA DaRosa. 2011. "Applying Multimedia Design Principles Enhances Learning in Medical Education." *Medical Education* 45(8):818-826. DOI: 10.1111/j.1365-2923.2011.03988.x.

Joung W, B Hesketh and A Neal. 2006. "Using "War Stories" to Train for Adaptive Performance: Is it Better to Learn from Error or Success?" *Applied Psychology* 55(2):282-302. DOI: 10.1111/j.1464-0597.2006.00244.x.

Kalyuga S, P Chandler and J Sweller. 1998. "Levels of Expertise and Instructional Design." *Human Factors* 40(1):1-17. DOI: 10.1518/001872098779480587.

Kapur M. 2012. "Productive Failure in Learning the Concept of Variance." *Instructional Science* 40(4):651-672. DOI: 10.1007/s11251-012-9209-6.

Kapur M. 2014. "Productive Failure in Learning Math." *Cognitive Science* 38(5):1008-1022. DOI: 10.1111/cogs.12107.

Kaye W and ME Mancini. 1986. "Retention of Cardiopulmonary Resuscitation Skills by Physicians, Registered Nurses, and the General Public." *Critical Care Medicine* 14(7):620-622.

Keith N and M Frese. 2008. "Effectiveness of Error Management Training: A Meta-Analysis." *Journal of Applied Psychology* 93(1):59-69. DOI: 10.1037/0021-9010.93.1.59.

Keller FS. 1968. ""Good-bye, Teacher . . ."." *Journal of Applied Behavior Analysis* 1(1):79-89. DOI: 10.1901/jaba.1968.1-79.

Keller FS. 1974. "Ten Years of Personalized Instruction." *Teaching of Psychology* 1(1):4-9. DOI: 10.1177/009862837400100102.

Kerfoot BP. 2009. "Learning Benefits of On-Line Spaced Education Persist for 2 Years." *The Journal of Urology* 181(6):2671-2673. DOI: 10.1016/j.juro.2009.02.024.

Kerfoot BP, HE Baker, MO Koch, D Connelly, DB Joseph and ML Ritchey. 2007. "Randomized, Controlled Trial of Spaced Education to Urology Residents in the United States and Canada." *The Journal of Urology* 177(4):1481-1487. DOI: 10.1016/j.juro.2006.11.074.

Kerfoot BP, Y Fu, H Baker, D Connelly, ML Ritchey and EM Genega. 2010a. "Online Spaced Education Generates Transfer and Improves Long-Term Retention of Diagnostic Skills: A Randomized Controlled Trial." *Journal of the American College of Surgeons* 211(3):331-337.e1. DOI: 10.1016/j.jamcollsurg.2010.04.023.

Kerfoot BP, EV Lawler, G Sokolovskaya, D Gagnon and PR Conlin. 2010b. "Durable Improvements in Prostate Cancer Screening from Online Spaced Education." *American Journal of Preventive Medicine* 39(5):472-478. DOI: 10.1016/j.amepre.2010.07.016.

Kirkbright S, J Finn, H Tohira, A Bremner, I Jacobs and A Celenza. 2014. "Audiovisual Feedback Device Use by Health Care Professionals During Cpr: A Systematic Review and Meta-Analysis of Randomised and Non-Randomised Trials." *Resuscitation* 85(4):460-471. DOI: 10.1016/j.resuscitation.2013.12.012.

Kirschner PA, J Sweller and RE Clark. 2006. "Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Teaching." *Educational Psychologist* 41(2):75-86. DOI: 10.1207/s15326985ep4102_1.

Kluger AN and A DeNisi. 1996. "The Effects of Feedback Interventions on Performance: A Historical Review, a Meta-Analysis, and a Preliminary Feedback Intervention Theory." *Psychological Bulletin* 119(2):254-284. DOI: 10.1037/0033-2909.119.2.254.

Knox AB. 1986. *Helping Adults Learn*, San Francisco: Jossey-Bass. ISBN 9781555420239.

Kraiger K and JK Ford. 2007. "The Expanding Role of Workplace Training: Themes and Trends Influencing Training Research and Practice." In *Historical Perspectives in Industrial and Organizational Psychology*, pp. 281-309. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Krampe RT and N Charness. 2018. "Aging and Expertise." In *The Cambridge Handbook of Expertise and Expert Performance, 2nd Edition*, pp. 835-856, eds: KA Ericsson, RR Hoffman, A Kozbelt and AM Williams. New York, NY: Cambridge University Press.

Lammers RL, MJ Byrwa, WD Fales and RA Hale. 2009. "Simulation-based Assessment of Paramedic Pediatric Resuscitation Skills." *Prehospital Emergency Care* 13(3):345-356. DOI: 10.1080/10903120802706161.

Larsen DP, AC Butler, AL Lawson and HL Roediger. 2013a. "The Importance of Seeing the Patient: Test-Enhanced Learning with Standardized Patients and Written Tests Improves Clinical Application of Knowledge." *Advances in Health Sciences Education* 18(3):409-425. DOI: 10.1007/s10459-012-9379-7.

Larsen DP, AC Butler and HL Roediger III. 2008. "Test-Enhanced Learning in Medical Education." *Medical Education* 42(10):959-966. DOI: 10.1111/j.1365-2923.2008.03124.x.

Larsen DP, AC Butler and HL Roediger III. 2009. "Repeated Testing Improves Long-Term Retention Relative to Repeated Study: A Randomised Controlled Trial." *Medical Education* 43(12):1174-1181. DOI: 10.1111/j.1365-2923.2009.03518.x.

Larsen DP, AC Butler and HL Roediger III. 2013b. "Comparative Effects of Test-Enhanced Learning and Self-Explanation on Long-Term Retention." *Medical Education* 47(7):674-682. DOI: 10.1111/medu.12141.

Latman NS and K Wooley. 1980. "Knowledge and Skill Retention of Emergency Care Attendants, EMT-As, and EMT-Ps." *Annals of Emergency Medicine* 9(4):183-189. DOI: 10.1016/S0196-0644(80)80003-5.

Levine AI, AD Schwartz, EO Bryson and S DeMaria Jr. 2012. "Role of Simulation in US Physician Licensure and Certification." *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine* 79(1):140-153. DOI: 10.1002/msj.21291.

Macchiarella ND, PK Arban and SM Doherty. 2006. "Transfer of Training from Flight Training Devices to Flight for ab-initio Pilots." *International Journal of Applied Aviation Studies* 6(2):299-314.

Mackay S, P Morgan, V Datta, A Chang and A Darzi. 2002. "Practice Distribution in Procedural Skills Training." *Surgical Endoscopy And Other Interventional Techniques* 16(6):957-961. DOI: 10.1007/s00464-001-9132-4.

Mayer R, J Heiser and S Lonn. 2001. "Cognitive Constraints on Multimedia Learning: When Presenting More Material Results in Less Understanding." *Journal of Educational Psychology* 93(1):187-198. DOI: 10.1037/0022-0663.93.1.187.

Mayer RE. 1989. "Systematic Thinking Fostered by Illustrations in Scientific Text." *Journal of Educational Psychology* 81(2):240-246. DOI: 10.1037/0022-0663.81.2.240.

Mayer RE. 2001. *Multi-Media Learning, 1st Edition*, NY: Cambridge University Press. ISBN 978-0521787499.

Mayer RE. 2004. "Should There Be a Three-Strikes Rule Against Pure Discovery Learning?" *American Psychologist* 59(1):14-19. DOI: 10.1037/0003-066X.59.1.14.

Mayer RE. 2008. "Applying the Science of Learning: Evidence-Based Principles for the Design of Multimedia Instruction." *American Psychologist* 63(8):760-769. DOI: 10.1037/0003-066X.63.8.760.

Mayer RE. 2011. *Applying the Science of Learning, 1st Edition*, San Francisco: Pearson. ISBN 978-0136117575.

Mayer RE. 2017. "Instruction Based on Visualizations." In *Handbook of Research on Learning and Teaching, 2nd Edition*, pp. 483-501, eds: RE Mayer and PA Alexander. NY: Routledge.

Mayer RE and PA Alexander, Eds. 2017. *Handbook of Research on Learning and Instruction, 2nd Edition*. Educational Psychology Handbook. NY: Routledge.

Mayer RE, A Stull, K DeLeeuw, K Almeroth, B Bimber, D Chun, M Bulger, J Campbell, A Knight and H Zhang. 2009. "Clickers in College Classrooms: Fostering Learning with Questioning Methods in Large Lecture Classes." *Contemporary Educational Psychology* 34(1):51-57. DOI: 10.1016/j.cedpsych.2008.04.002.

McKenna SP and AI Glendon. 1985. "Occupational First Aid Training: Decay in Cardiopulmonary Resuscitation (CPR) Skills." *Journal of Occupational Psychology* 58(2):109-117. DOI: 10.1111/j.2044-8325.1985.tb00186.x.

McKinney Jr. EH and KJ Davis. 2003. "Effects of Deliberate Practice on Crisis Decision Performance." *Human Factors* 45(3):436-444.

McLean GMT, S Lambeth and T Mavin. 2016. "The Use of Simulation in Ab Initio Pilot Training." *The International Journal of Aviation Psychology* 26(1-2):36-45. DOI: 10.1080/10508414.2016.1235364.

Melton AW. 1959. "The Science of Learning and the Technology of Educational Methods." *Harvard Educational Review* 29:96-106.

Menekse M, GS Stump, S Krause and MTH Chi. 2013. "Differentiated Overt Learning Activities for Effective Instruction in Engineering Classrooms." *Journal of Engineering Education* 102(3):346-374. DOI: 10.1002/jee.20021.

Mengelkoch RF, JA Adams and CA Gainer. 1971. "The Forgetting of Instrument Flying Skills." *Human Factors* 13(5):397-405.

Molesworth BRC, L Bennett and EJ Kehoe. 2011. "Promoting Learning, Memory, and Transfer in a Time-Constrained, High Hazard Environment." *Accident Analysis & Prevention* 43(3):932-938. DOI: 10.1016/j.aap.2010.11.016.

Moulton C-AE, A Dubrowski, H MacRae, B Graham, E Grober and R Reznick. 2006. "Teaching Surgical Skills: What Kind of Practice Makes Perfect?: A Randomized, Controlled Trial." *Annals of Surgery* 244(3):400-409. DOI: 10.1097/01.sla.0000234808.85789.6a.

Newell A and PS Rosenbloom. 1981. "Mechanisms of Skill Acquisition and the Law of Practice." In *Cognitive Skills and Their Acquisition*, pp. 1-55, ed: JR Anderson. Hillsdale, NJ: Lawrence Erlbaum Associates.

Niles D, RM Sutton, A Donoghue, MS Kalsi, K Roberts, L Boyle, A Nishisaki, KB Arbogast, M Helfaer and V Nadkarni. 2009. "“Rolling Refreshers”: A Novel Approach to Maintain CPR Psychomotor Skill Competence." *Resuscitation* 80(8):909-912. DOI: 10.1016/j.resuscitation.2009.04.021.

Nodine CF, HL Kundel, C Mello-Thoms, SP Weinstein, SG Orel, DC Sullivan and EF Conant. 1999. "How Experience and Training Influence Mammography Expertise." *Academic Radiology* 6(10):575-585. DOI: 10.1016/S1076-6332(99)80252-9.

Palinscar AS and AL Brown. 1984. "Reciprocal Teaching of Comprehension-Fostering and Comprehension-Monitoring Activities." *Cognition and Instruction* 1(2):117-175. DOI: 10.1207/s1532690xci0102_1.

Palter VN and TP Grantcharov. 2014. "Individualized Deliberate Practice on a Virtual Reality Simulator Improves Technical Performance of Surgical Novices in the Operating Room: A Randomized Controlled Trial." *Annals of Surgery* 259(3):443-448. DOI: 10.1097/sla.0000000000000254.

Pintrich PR, DR Cross, RB Kozma and WJ McKeachie. 1986. "Instructional Psychology." *Annual Review of Psychology* 37(1):611-651. DOI: 10.1146/annurev.ps.37.020186.003143.

Prophet WW. 1976. *Long-Term Retention of Flying Skills: A Review of the Literature*. Report HumRRO-FR-ED(P)-76-35. Alexandria, VA: Human Resources Research Organization.

Renkl A. 2014. "Toward an Instructionally Oriented Theory of Example-Based Learning." *Cognitive Science* 38(1):1-37. DOI: 10.1111/cogs.12086.

Renkl A. 2017. "Instruction Based on Examples." In *Handbook of Research on Learning and Teaching, 2nd Edition*, pp. 325-348, eds: RE Mayer and PA Alexander. NY: Routledge.

Rittle-Johnson B and AM Loehr. 2017. "Instruction Based on Self-Explanation." In *Handbook of Research on Learning and Teaching, 2nd Edition*, pp. 349-364, eds: RE Mayer and PA Alexander. NY: Routledge.

Roberts KE, RL Bell and AJ Duffy. 2006. "Evolution of Surgical Skills Training." *World Journal of Gastroenterology* 12(20):3219-3224. DOI: 10.3748/wjg.v12.i20.3219.

Roediger HL. 2013. "Applying Cognitive Psychology to Education:Translational Educational Science." *Psychological Science in the Public Interest* 14(1):1-3. DOI: 10.1177/1529100612454415.

Roscoe SN. 1971. *Incremental Transfer Effectiveness*. Technical Report ARL-70-5/AFOSR-70-1. Savoy, IL: University of Illinois at Urbana-Champaign, Aviation Research Laboratory.

Rose AM, MY Czarnolewski, FE Gragg, SH Austin, P Ford, J Doyle and JD Hagman. 1985. *Acquisition and Retention of Soldiering Skills*. Technical Paper 671. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Rosen MA, E Salas, D Pavlas, R Jensen, D Fu and D Lampton. 2010. "Demonstration-Based Training: A Review of Instructional Features." *Human Factors* 52(5):596-609. DOI: 10.1177/0018720810381071.

Rosenshine B and C Meister. 1994. "Reciprocal Teaching: A Review of the Research." *Review of Educational Research* 64(4):479-530. DOI: 10.3102/00346543064004479.

Sabol MA and RA Wisher. 2001. "Retention and Reacquisition of Military Skills." *Military Operations Research* 6(1):59-80.

Salas E and JA Cannon-Bowers. 2001. "The Science of Training: A Decade of Progress." *Annual Review of Psychology* 52(1):471-499. DOI: 10.1146/annurev.psych.52.1.471.

Sanquist TF, SL Morrow, AL D'Agostino, NM Hughes and CM Franklin. 2018. *Human Factors in Nondestructive Examination: Manual Ultrasonic Testing Task Analysis and Field Research*. PNNL-27441. Richland, WA: Pacific Northwest National Laboratory. ADAMS Accession No. ML18176A055.

Sawyer JE, R Obeid, D Bublitz, AM Schwartz, PJ Brooks and AS Richmond. 2017. "Which Forms of Active Learning Are Most Effective: Cooperative Learning, Writing-to-Learn, Multimedia Instruction, or Some Combination?" *Scholarship of Teaching and Learning in Psychology* 3(4):257-271. DOI: 10.1037/stl0000095.

Schendel JD, JL Shields and MS Katz. 1978. *Retention of Motor Skills: Review*. Technical Paper 313. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Semb GB, MA Fitch, JA Ellis and MB Kuti. 2000. "On-the-Job Training (OJT): Theory, Research and Practice." In *Training & Retraining: A Handbook for Business, Industry, Government, and the Military*, pp. 289-311, eds: S Tobias and JD Fletcher. New York: Macmillan Library Reference USA.

Seymour NE, AG Gallagher, SA Roman, MK O'Brien, VK Bansal, DK Andersen and RM Satava. 2002. "Virtual Reality Training Improves Operating Room Performance: Results of a Randomized, Double-Blinded Study." *Annals of Surgery* 236(4):458-464.

Shields JL, SL Goldberg and JD Dressel. 1979. *Retention of Basic Soldiering Skills*. Technical Paper 1225. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Skinner BF. 1958. "Teaching Machines." *Science* 128(3330):969-977. DOI: 10.1126/science.128.3330.969.

Sonnentag S and BM Kleine. 2000. "Deliberate Practice at Work: A Study with Insurance Agents." *Journal of Occupational and Organizational Psychology* 73(1):87-102. DOI: 10.1348/096317900166895.

Spitzer HF. 1939. "Studies in Retention." *Journal of Educational Psychology* 30(9):641-656. DOI: 10.1037/h0063404.

Spruit EN, GPH Band and JF Hamming. 2015. "Increasing Efficiency of Surgical Training: Effects of Spacing Practice on Skill Acquisition and Retention in Laparoscopy Training." *Surgical Endoscopy* 29(8):2235-2243. DOI: 10.1007/s00464-014-3931-x.

Stewart JE, JA Dohme and RT Nullmeyer. 2002. "U.S. Army Initial Entry Rotary-Wing Transfer of Training Research." *The International Journal of Aviation Psychology* 12(4):359-375. DOI: 10.1207/S15327108IJAP1204_3.

Swinnen SP, RA Schmidt, DE Nicholson and DC Shapiro. 1990. "Information Feedback for Skill Acquisition: Instantaneous Knowledge of Results Degrades Learning." *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16(4):706-716. DOI: 10.1037/0278-7393.16.4.706.

Taylor PJ, DF Russ-Eft and DWL Chan. 2005. "A Meta-Analytic Review of Behavior Modeling Training." *Journal of Applied Psychology* 90(4):692-709. DOI: 10.1037/0021-9010.90.4.692.

Tolsgaard MG, C Ringsted, E Dreisler, LN Nørgaard, JH Petersen, ME Madsen, NLC Freiesleben, JL Sørensen and A Tabor. 2015. "Sustained Effect of Simulation-Based Ultrasound Training on Clinical Performance: A Randomized Trial." *Ultrasound in Obstetrics & Gynecology* 46(3):312-318. DOI: 10.1002/uog.14780.

van Dongen K, P Mitra, M Schijven and I Broeders. 2011. "Distributed Versus Massed Training: Efficiency of Training Psychomotor Skills." *Surgical Techniques Development* 1(1):e17. DOI: 10.4081/std.2011.e17.

Weaver FJ, AG Ramirez, SB Dorfman and AE Raizner. 1979. "Trainees' Retention of Cardiopulmonary Resuscitation: How Quickly They Forget." *JAMA* 241(9):901-903. DOI: 10.1001/jama.1979.03290350021013.

Welsher A and LEM Grierson. 2017. "Enhancing Technical Skill Learning through Interleaved Mixed-Model Observational Practice." *Advances in Health Sciences Education* 22(5):1201-1211. DOI: 10.1007/s10459-017-9759-0.

Westermann K and N Rummel. 2012. "Delaying Instruction: Evidence from a Study in a University Relearning Setting." *Instructional Science* 40(4):673-689. DOI: 10.1007/s11251-012-9207-8.

Wik L, J Kramer-Johansen, H Myklebust, H Sørebø, L Svensson, B Fellows and PA Steen. 2005. "Quality of Cardiopulmonary Resuscitation During Out-of-Hospital Cardiac Arrest." *JAMA* 293(3):299-304. DOI: 10.1001/jama.293.3.299.

Willis RE and KR Van Sickle. 2015. "Current Status of Simulation-Based Training in Graduate Medical Education." *Surgical Clinics of North America* 95(4):767-779. DOI: 10.1016/j.suc.2015.04.009.

Wood D, JS Bruner and G Ross. 1976. "The Role of Tutoring in Problem Solving." *Journal of Child Psychology and Psychiatry* 17(2):89-100. DOI: 10.1111/j.1469-7610.1976.tb00381.x.

Wright RH. 1973. *Retention of Flying Skills and Refresher Training Requirements: Effects of Nonflying and Proficiency Flying*. Report HumRRO-TR-73-32. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

PNNL-29761

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov